Universidade de Aveiro

Departamento de Electrónica e Telecomunicações

Codificação de fala por modelos variáveis no tempo

Paulo Duarte Ferreira Gouveia

Lic. em Engª Electrónica e Telecomunicações

pela

Universidade de Aveiro

Abril de 1996

Codificação de fala
por
modelos variáveis no tempo

Paulo Duarte Ferreira Gouveia

Lic. em Engª Electrónica e Telecomunicações

pela

Universidade de Aveiro

Dissertação submetida para satisfação parcial
dos requisitos do programa de Mestrado em
Engenharia Electrónica e Telecomunicações

Universidade de Aveiro

Departamento de Electrónica e Telecomunicações

Abril de 1996

Tese realizada sob supervisão de

Prof. Dr. **Francisco António Cardoso Vaz**

Professor Catedrático do

Departamento de Electrónica e Telecomunicações

Universidade de Aveiro

Resumo

O trabalho apresentado nesta tese representa uma contribuição para a optimização da codificação da fala. Utilizam-se para o efeito modelos de codificação baseados em filtros LP (filtros de Predição Linear) de parâmetros variáveis no tempo, contrastando com os modelos fixos utilizados nos métodos convencionais. Nestes, a adaptação dos filtros de predição realiza-se simplesmente através de actualizações periódicas dos seus parâmetros, não traduzindo por isso uma evolução gradual e contínua ao longo do tempo.

A técnica utilizada na implementação dos modelos variáveis tem por base a utilização de funções do tipo *B-spline* na representação das formas de onda dos parâmetros LP.

Para o estudo da viabilidade do modelo proposto, analisou-se o desempenho de um *vocoder* de predição linear incluindo, quer o modelo LP de parâmetros variáveis, quer o modelo LP de parâmetros fixos convencional, por forma a possibilitar a comparação de desempenhos.

Dos resultados obtidos concluímos que a codificação de fala por modelos variáveis no tempo, embora não tenha evidenciado vantagens convincentes, pode ser encarada como outra forma de codificação, competindo por isso com as metodologias já existentes.

Abstract

The work presented in this thesis aims at to be a contribution to speech coding. To accomplish this objective, coding models based on LP filters (Linear Predictive Filters) with time-varying parameters are used, and compared with fixed models used in conventional methods. In these models, the predictive filters adaptation is carried on simply through periodic updatings of its parameters, therefore doesn't representing a gradual and continuous evolution in time.

The technique used in varying models implementation is based on the utilization of B-spline like functions to represent the LP parameters waveforms.

In order to make a viability study of the proposed model, the performance of a linear predictive vocoder was analyzed, including both the LP model with varying parameters and the conventional LP model with fixed parameters, thus enabling the comparison of their performances.

From the results, we concluded that speech coding by time-varying models, although it had not demonstrated clear benefits, can be viewed as another coding way, therefore competing with the already existing methodologies.

Agradecimentos

Infelizmente não me é possível agradecer toda a colaboração concedida por todos aqueles que contribuíram com informação e tempo para a compilação deste trabalho. Porém, os seus esforços são por mim profundamente reconhecidos.

Um agradecimento especial ao Prof. Dr. Francisco António Cardoso Vaz pela orientação, incentivo e motivação que me soube incutir, sem os quais não seria, de certo, possível a conclusão deste trabalho.

Um agradecimento também muito especial à Prudência pela sua compreensão, dedicação e apoio, extremamente valiosos, particularmente em horas de maior frustração vividas ao longo deste trabalho.

Ao Instituto Superior Politécnico de Bragança pelas condições de trabalho que me proporcionou na fase final da dissertação.

À JNICT pelo apoio financeiro prestado durante a realização do curso de Mestrado.

Índice de Conteúdos

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A investigação em processamento digital de sinal tem vindo a assumir um papel extremamente importante particularmente na área da codificação da fala, pois representa o processo privilegiado no sentido de permitir uma eficiente utilização de recursos, quer de transmissão, quer de armazenamento de sinais de voz no seu formato digital.

Os principais algoritmos de codificação hoje adoptados têm por base um modelo digital deduzido directamente a partir de um modelo físico representativo do sistema humano de produção de fala. Como na metodologia convencional o modelo digital é adaptado, não de uma forma gradual e contínua, à semelhança do que acontece no sistema humano, mas através de actualizações periódicas, parece preferível utilizar-se um modelo de parâmetros variáveis no tempo, de forma a que o modelo assim obtido reproduza mais fielmente o comportamento do sistema modelado. Para o efeito, pensou-se em utilizar funções *B-spline* na representação dos parâmetros do modelo. Embora estas funções sejam já utilizadas com sucesso noutras áreas, como por exemplo, utilizadas como funções interpoladoras em computação gráfica, a sua aplicabilidade ainda não foi devidamente explorada em processamento de voz. É pois nosso objectivo estudar a aplicabilidade dessas funções na área da codificação de fala.

1.2 Meios Utilizados

Para a simulação e avaliação do modelo proposto desenvolveram-se conjuntos de rotinas em *Matlab*, versão 4.0 para o *Windows*, não apenas as de suporte directo à

implementação dos algoritmos de codificação/descodificação, como também, rotinas responsáveis pela visualização dos resultados.

A simulação foi implementada num PC 486DX2 50Mz, e recorreu-se a um outro, munido de uma placa de som *Sound Blaster* para possibilitar a comparação perceptual dos resultados obtidos pelos diferentes métodos.

1.3 Notação Adoptada

Se nada for referido em contrário, a notação geral adoptada neste trabalho é a seguinte: letras minúsculas denotarão sinais no domínio do tempo e maiúsculas sinais no domínio da frequência; letras a cheio (*bold*) serão utilizadas na representação de matrizes e vectores.

Assim, $s(n)$ representará um sinal discreto no tempo, onde n é um índice inteiro representando o número da amostra. Este sinal discreto estará relacionado com o correspondente sinal analógico, $s_a(t)$, por $s(n) = s_a(t)\big|_{t=nT} = s_a(nT)$, sendo T o período de amostragem.

1.4 Estrutura da Tese

Esta tese encontra-se organizada por capítulos, que por sua vez estão subdivididos em secções, que poderão ainda subdividir-se em subsecções. Em traços gerais encontra-se organizada como se segue.

No capítulo 2 é apresentada uma descrição fonética, uma vez que o conhecimento sobre a diversidade de sons que constituem a fala é indispensável ao desenvolvimento e compreensão de técnicas de codificação específicas da fala.

O capítulo 3 é dedicado à descrição do mecanismo humano de produção de fala, pois as técnicas utilizadas na codificação exploram, por norma, os princípios acústicos relacionados com o processo natural de produção de fala.

No capítulo 4 demonstra-se de que forma é que os conceitos acústicos nos podem conduzir ao desenvolvimento do modelo digital de produção de fala mais adoptado pelos codificadores existentes.

No capítulo 5 é feito um levantamento sobre os vários algoritmos convencionais utilizados na codificação da fala.

No capítulo 6 introduzimos a estratégia a seguir na implementação de um modelo variável no tempo a partir de funções do tipo *B-spline*.

O capítulo 7 refere-se às implementações realizadas para avaliação da viabilidade da metodologia proposta, e inclui a consequente discussão de resultados.

No capítulo 8, em jeito de conclusão, são feitos alguns comentários finais, e apontam-se algumas orientações sobre um possível trabalho a desenvolver futuramente.

Por fim, em apêndice, é feita uma apresentação dos fundamentos teóricos associados às funções *B-spline*, bem como os relacionados com a formalização de curvas a partir dessas funções ou de outras da mesma família, necessários ao desenvolvimento dos modelos variáveis baseados nessas funções.

Capítulo 2

Fonética

A voz é um sinal não estacionário formado por uma sequência de sons articulados. Nesta cadeia sonora, são os sons e as transições entre eles que servem de símbolos de representação da informação [Rabiner (79)]. Os sons pronunciados numa palavra têm funções distintas no funcionamento de uma língua:

♦ Alguns são os responsáveis pela diferenciação de palavras, que no restante sejam coincidentes — todo o som capaz de, por si só, estabelecer distinção entre duas palavras, representa um *fonema*. É o caso, por exemplo, da vogal tónica nas palavras ***e**rro* e *almo**ç**o*, onde a diferença de timbre utilizado — fechado ou aberto — é suficiente para as identificar, quer como substantivos, quer como verbos. Podemos referir igualmente as palavras ***d**ia*, ***f**ia*, ***m**ia*, ***p**ia*, ***t**ia* e ***v**ia* como exemplos ilustrativos deste tipo de sons, pois a única particularidade que as distingue reside no som consonântico inicial.

♦ Outros poderão traduzir apenas distintas pronúncias de um mesmo fonema. Essas variantes de pronúncia, que não impedem a identificação da palavra em causa, podem ser de natureza regional, social ou, até mesmo, individual.

Numa descrição fonética deveria considerar-se como os sons são produzidos e como são percebidos. É através da percepção auditiva que percebemos a variedade de sons e a sua funcionalidade. Contudo, embora a descrição fonética com base na percepção — fonética acústica — não seja de somenos importância do que a que se baseia na produção dos sons — fonética fisiológica —, os fonemas são descritos e classificados fundamentalmente em função dos movimentos dos órgãos que participam na sua formação. A razão de ser desta realidade deve-se ao facto da fonética acústica ainda se encontrar no seu começo. A descrição acústica de um

fonema ainda não se faz com termos precisos, semelhantes aos que se utilizam para descrever os movimentos dos órgãos responsáveis pela sua produção. A fonética fisiológica é uma especialidade que se encontra já numa fase de desenvolvimento bastante avançada. Esta é, portanto, a razão pela qual a fonologia é, em geral, descrita com base nas características articulatórias.

De forma a podermos simbolizar na escrita a pronúncia real de um som, torna-se necessário a utilização de um alfabeto especial — o alfabeto fonético — que, contrariamente ao que acontece nos alfabetos convencionais, cada símbolo traduza um único som. Daqui para a frente, adoptaremos o alfabeto fonético da Tabela 2-1 sempre que houver necessidade de representar fonemas por escrito. Estes transcrevem-se sempre entre duas barras obliquas, / /.

VOGAIS

Vogais ORAIS	*Exemplificação*
/i/	l***i***vro
/ê/	P***e***dro
/é/	t***e***rra
/á/	p***a***to
/a/	jog***a***
/ó/	g***o***la
/ô/	p***o***ço
/u/	p***u***lar
/e/	s***e***car

Vogais NASAIS	*Exemplificação*
/$\tilde{i}$/	p***in***to
/$\tilde{e}$/	d***en***te
/ã/	c***an***to
/õ/	p***on***te
/$\tilde{u}$/	f***un***do

SEMIVOGAIS

Semivogais	*Exemplificação*
/y/	pa***i***
/w/	pa***u***

CONSOANTES

Cons. ORAIS	*Exemplificação*
/p/	***p***á
/b/	***b***ata
/t/	***t***arde
/d/	***d***ado
/k/	***c***ão
/g/	***g***ato
/s/	***s***ábado
/z/	cas***s***a
/x/	***ch***ão
/j/	***j***ardim
/f/	***f***ado
/v/	***v***aca
/l/	***l***ado
/lh/	fi***lh***o
/r/	po***r***ta
/rr/	ca***rr***o
/R/	po***r***ta (*velar*)
/RR/	ca***rr***o (*velar*)

Cons. NASAIS	*Exemplificação*
/m/	***m***ãe
/n/	***n***ada
/nh/	pi***nh***o

Tabela 2-1: Alfabeto fonético português.

2.1 Classificação Fonética

Os fonemas classificam-se em:

♦ *VOGAIS* – do ponto de vista articulatório, as vogais representam sons produzidos pelas cordas vocais e modulados pela forma do tracto vocal. Durante a produção deste tipo de sons, o tubo acústico formado pelo tracto vocal permanece aberto e sempre com a mesma configuração.

♦ *CONSOANTES* – ao contrário das vogais, na produção das consoantes existem sempre na cavidade bucal obstáculos à passagem do ar.

♦ *SEMIVOGAIS* – entre as vogais e as consoantes encontram-se as semivogais. Estes fonemas são os sons associados ao *i* ou ao *u* quando, juntos a uma vogal, formam com ela uma só silaba (ditongo). Assim, o *i* em *vário* (*vá-rio*) traduz a semivogal /y/, mas em *rio* (*ri-o*) traduz a vogal /i/. De igual forma o *u* simboliza uma semivogal, por exemplo em *chapéu* (*cha-péu*), mas uma vogal em *rua* (*ru-a*).

2.1.1 Vogais

As vogais são produzidas com uma fonte sonora quase periódica ao nível da glote e sem obstruções significativas à passagem do ar pelo tracto vocal. As anti-ressonâncias não são significativas e o amortecimento das formantes é pequeno, apresentando por isso, larguras de banda relativamente estreitas. A energia das formantes de baixa frequência é superior à das formantes de alta frequência.

As vogais podem ser classificadas:

- quanto à zona de articulação;
- quanto ao grau de elevação da língua;
- quanto ao timbre;
- quanto à intensidade; e
- quanto ao papel cavidade nasal.

❍ Zona de articulação

Embora as vogais sejam pronunciadas com a via bucal livre, a posição de certos órgãos articuladores é determinante para a produção de vogais distintas. Basta a língua elevar-se progressivamente em direcção ao palato duro ou ao véu palatino (palato mole) para se obter uma série de vogais distintas, tal como ilustrado na Figura 2-1.

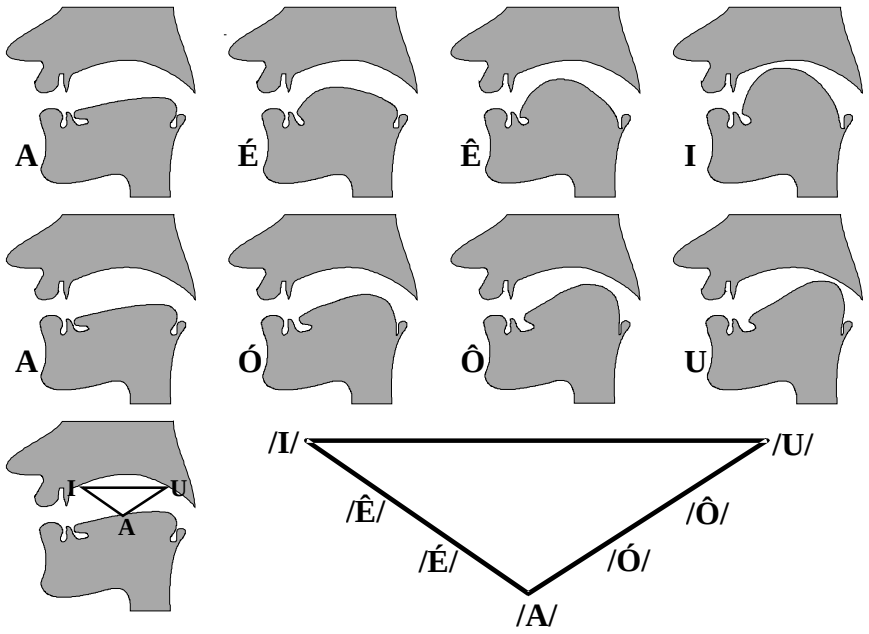


Figura 2-1: Zona de articulação das vogais.

Quanto à zona de articulação, as vogais são classificadas em:

- *ANTERIORES*;
- *MÉDIAS*;
- *POSTERIORES*.

Como se depreende da Figura 2-1, as vogais /é/, /ê/, /i/ são produzidas com a parte anterior da língua elevada em direcção ao palato duro, sendo por isso designadas por vogais *anteriores* (ou *palatais*). Já as vogais /ó/, /ô/, /u/ são produzidas com a parte posterior da língua, e elevada em direcção ao véu palatino, justificando dessa forma a designação de vogais *posteriores* (ou *velares*). Entre estas, situa-se a vogal /a/ que, em virtude do seu modo de articulação — com a língua praticamente em repouso —, denomina-se vogal *média*.

❍ **Elevação da língua**

Como ilustrado ainda na Figura 2-1, as vogais são articuladas com diferentes elevações da língua. Enquanto que as vogais /i/ e /u/ são pronunciadas com uma acentuada elevação da língua, na vogal /a/ a mesma encontra-se quase em repouso. Daí, as primeiras serem classificadas como vogais *altas*, e a última como vogal *baixa*. As restantes vogais — /é/, /ê/, /ó/, /ô/ —, por serem pronunciadas com elevações intermédias da língua, são consideradas vogais *mediais*.

❍ **Timbre**

O timbre é a principal característica que distingue as vogais. Resulta da composição do tom fundamental — imposto pelas cordas vocais — com os seus harmónicos. Este tipo de composição deriva da modulação imposta pela forma do tracto vocal na onda acústica emitida pela laringe.

As vogais dizem-se abertas, fechadas ou reduzidas, consoante o grau de abertura do tubo de ressonância — cavidade bucal — durante a produção das mesmas. As vogais são de timbre:

♦ *ABERTO* – quando produzidas com o tubo de ressonância com uma abertura considerável (é o caso das vogais /é/, /á/, /ó/);

♦ *FECHADO* – quando produzidas através de uma estreita abertura do tubo de ressonância (é o caso das vogais /i/, /ê/, /u/, /ô/).

❍ **Intensidade**

A intensidade está relacionada com a força expiratória com que se pronunciam os fonemas. É ela a principal característica que distingue a sílaba tónica da sílaba átona.

Uma vez que as sílabas tónicas são pronunciadas com maior intensidade, a diferença entre timbre aberto e fechado faz-se mais sentir em posição tónica do que em posição átona; por isso, a distinção entre /ê/ e /é/, e entre /ô/ e /ó/ anula-se em sílaba átona.

❍ Vogais orais e vogais nasais

Até agora foram mencionadas sete vogais *orais*, pois todas elas eram produzidas com o véu palatino elevado contra a parede posterior da faringe, impedindo dessa forma a passagem do ar pela cavidade nasal. No entanto, além dessas, a língua portuguesa possui cinco vogais *nasais*, as quais, são pronunciadas com o véu palatino abaixado, de modo a permitir que parte da corrente expiratória passe pelas cavidades nasais — a restante continua a fluir pela cavidade bucal —, e assim introduzir alguma modulação nasal no som emitido.

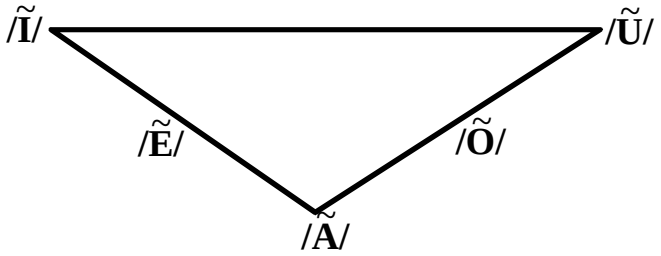


Figura 2-2: Vogais nasais.

Na língua portuguesa, todas as vogais nasais são fechadas. Obviamente, a nasalidade de uma vogal numa palavra pode ser utilizada para a diferenciar de outras palavras desprovidas dessa nasalidade. É o que acontece, por exemplo, entre as palavras *lã* e *lá*, *senda* e *seda*, e entre *mundo* e *mudo*.

Recapitulando, a Tabela 2-2 resume todos os critérios de classificação das vogais.

Zona de articulação		ANTERIORES		MÉDIAS		POSTERIORES	
Papel das cavidades bucal e nasal		ORAIS	NASAIS	ORAIS	NASAIS	ORAIS	NASAIS
Elevação da língua	Timbre						
ALTAS	FECHADAS	/i/	/ĩ/			/u/	/ũ/
	REDUZIDAS	/e/	/ĩ/			/o/	/ũ/
MEDIAIS	FECHADAS	/ê/	/ẽ/			/ô/	/õ/
	ABERTAS	/é/				/ó/	
BAIXAS	FECHADAS				/ã/		
	ABERTAS			/a/			
	REDUZIDAS			/á/	/ã/		

Tabela 2-2: Classificação das vogais.

2.1.2 Consoantes

As consoantes são produzidas com uma obstrução significativa ao nível da cavidade bucal, e caracterizam-se pela presença de anti-ressonâncias que afectam todo o espectro. Desse modo, as formantes possuem larguras de banda consideravelmente maiores do que as dos sons não consonânticos, e a energia global é também francamente inferior.

De uma forma análoga ao que sucedia nas vogais, as consoantes podem ser classificadas em função de quatro critérios, de base essencialmente articulatória:

- quanto ao modo de articulação;
- quanto à zona de articulação;
- quanto ao papel das cordas vocais; e
- quanto ao papel da cavidade nasal.

❍ **Modo de articulação**

Contrariamente ao que sucede na produção das vogais — geradas com a passagem do ar através da cavidade bucal livre —, na articulação das consoantes a corrente expiratória sofre algum tipo de perturbação, introduzida por uma obstrução numa determinada parte da boca, que a interrompe momentaneamente ou a comprime parcialmente. Consoante se obstrua total ou parcialmente a cavidade bucal, as consoantes resultantes dizem-se, respectivamente, *oclusivas* ou *constritivas*.

♦ Consoantes OCLUSIVAS

As consoantes oclusivos (ou plosivas) são produzidas por obstrução completa duma zona da cavidade bucal, criando uma elevada pressão de ar nessa posição, e libertando-a de seguida abruptamente. São oclusivas as consoantes:

/b/ /d/ /g/ /k/ /p/ /t/

♦ Consoantes CONSTRITIVAS

Ainda quanto ao modo de articulação, podem-se distinguir três tipos de consoantes constritivas:

- as FRICATIVAS – provocadas pela constrição dum determinado ponto da cavidade bucal, provocando dessa forma a passagem turbulenta do ar nesse ponto. É o caso das consoantes:

  /f/ /j/ /s/ /v/ /x/ /z/

- as LATERAIS – caracterizadas pela passagem do fluxo de ar pelos lados da cavidade bucal, em virtude da parte central da boca se encontrar obstruída pela língua, que se encontra em contacto com o palato ou com os alvéolos dos dentes. São laterais as consoantes:

  /l/ /lh/

- e as VIBRANTES – caracterizadas pelo movimento vibratório rápido dum órgão elástico — língua ou véu palatino —, provocando brevíssimas oclusões da corrente expiratória. São vibrantes as consoantes:

  /r/ /rr/ /R/ /RR/

Os sons consonânticos podem ser também classificados em *contínuos* e *descontínuos*. Os sons descontínuos ou interrompidos caracterizam-se por um "ataque" abrupto a que estão associadas alterações bruscas das características espectrais. A estes, opõem-se os sons contínuos em que o "ataque" é gradual. São consideradas contínuas as consoantes fricativas e as laterais, e interrompidas as oclusivas e as vibrantes.

❍ **Zona de articulação**

Embora as consoantes pronunciadas dependam do tipo de constrição utilizada, a localização dessa obstrução a nível da cavidade bucal é igualmente determinante no tipo de som emitido. Consoante a posição da articulação, as consoantes classificam-se em:

♦ BILABIAIS – quando articuladas através do contacto dos lábios:

/b/ /m/ /p/

♦ LABIODENTAIS – quando resultam da constrição do ar pela aproximação entre o lábio inferior e os dentes incisivos superiores:

/f/ /v/

♦ LINGUODENTAIS – quando articuladas por intermédio do contacto da ponta da língua com a face interna dos dentes superiores:

/d/ /n/ /t/

♦ ALVEOLARES – quando articuladas com a língua elevada em direcção aos alvéolos superiores dos dentes:

/l/ /r/ /rr/ /s/ /z/

♦ PALATAIS – quando obtidas pela aproximação ou contacto do dorso da língua com o palato duro:

/j/ /x/ /lh/ /nh/

♦ VELARES – quando articuladas através da aproximação ou do contacto da parte posterior da língua com o véu palatino:

/g/ /k/ /R/ /RR/

❍ **Papel das cordas vocais**

Enquanto que as vogais são sempre vozeadas, as consoantes podem-no ser ou não.

♦ São VOZEADAS as consoantes:

/b/ /d/ /g/ /j/ /l/ /lh/ /m/
/n/ /nh/ /r/ /rr/ /R/ /RR/ /v/
/z/

♦ As restantes são NÃO VOZEADAS:

/f/ /p/ /k/ /s/ /t/ /x/

A vibração ou não das cordas vocais aquando da fonação é suficiente para diferenciar algumas das consoantes na língua portuguesa. Os exemplos ilustrativos desta constatação são apresentados na Tabela 2-3.

		NÃO VOZEADAS	**VOZEADAS**	EXEMPLIFICAÇÃO
OCLUSIVAS	BILABIAIS	/p/	/b/	***p**ato* ***b**ato*
	LINGUODENTAIS	/t/	/d/	***t**acto* ***d**ato*
	VELARES	/k/	/g/	***c**ato* ***g**ato*
FRICATIVAS	LABIODENTAIS	/f/	/v/	***f**aca* ***v**aca*
	ALVEOLARES	/s/	/z/	***s**elo* ***z**elo*
	PALATAIS	/x/	/j/	***ch**ato* ***j**acto*

Tabela 2-3: Vozeamento das consoantes.

❍ Papel da cavidade nasal

Tal como as vogais, as consoantes podem ser orais ou nasais. Orais, quando a corrente expiratória passa apenas pelo canal bucal; nasais, quando parte da corrente flui pela cavidade nasal.

O acoplamento do tracto nasal durante a articulação das consoantes manifesta-se pela presença de um som nasal que se caracteriza pela presença de duas formantes estáveis a cerca de 200 Hz e 2500 Hz.

São nasais as consoantes:

/m/ /n/ /nh/

Todas as outras são consoantes orais.

A Tabela 2-4 apresenta de uma forma resumida a classificação de todas as consoantes portuguesas, segundo critérios de base articulatória.

papel cavidades bucal e nasal	ORAIS							NASAIS
modo de articulação	OCLUSIVAS		CONSTRITIVAS					
			FRICATIVAS		LATERAIS	VIBRANTES		
						SIMPLES	MÚLTIPLA	
papel cordas vocais	vozeadas	não voz.	vozeadas	não voz.	vozeadas	vozeadas	vozeadas	vozeadas
zona de articulação								
BILABIAIS	/b/	/p/						/m/
LABIODENTAIS			/v/	/f/				
LINGUODENTAIS	/d/	/t/						/n/
ALVEOLARES			/z/	/s/	/l/	/r/	/rr/	
PALATAIS			/j/	/x/	/lh/			/nh/
VELARES	/g/	/k/				/R/	/RR/	

Tabela 2-4: Classificação fonética das consoantes.

2.2 Encontros Vocálicos e Consonânticos

❍ Ditongos

Quando juntas, uma vogal com uma semivogal formam um ditongo, que pode ser crescente ou decrescente, consoante a semivogal apareça, respectivamente, antes ou depois da vogal. Contudo, na língua portuguesa apenas os decrescentes são ditongos estáveis.

DITONGOS ORAIS

Ditongo	*Exemplificação*
/áy/	p***ai***
/áw/	m***au***
/éy/	r***éi***s
/êw/	m***eu***
/éw/	c***éu***
/iw/	v***iu***
/ôy/	b***oi***
/óy/	her***ói***
/ôw/	v***ou***
/uy/	az***ui***s

DITONGOS NASAIS

Ditongo	*Exemplificação*
/ãy/	m***ãe***, c***ãi***bra
/ãw/	m***ão***, vej***am***
/ ẽy /	v***em***, b***en***zinho
/õy/	p***õe***, serm***õe***s
/ ũy /	m***ui***, m***ui***to

Tabela 2-5: Ditongos decrescentes.

Os ditongos podem também ser orais ou nasais. Orais, quando a vogal que os formam é oral; nasais, quando a vogal é nasal. A Tabela 2-5 inclui todos os ditongos decrescentes da língua portuguesa.

❍ Hiatos

Por fim, ao encontro de duas vogais dá-se o nome de hiato. É o que acontece, por exemplo, na palavra *caí* , visto o encontro /*ai*/ soar em duas sílabas (*ca-í*). Pelo contrário, na palavra *cai* o *'a'* e o *'i'* soam na mesma sílaba, tratando-se por isso de um ditongo.

Embora existam encontros vocálicos absolutamente estáveis, outros existem porém com um comportamento algo instável. Assim, a palavra *lua* possuirá sempre

duas sílabas, ao passo que a palavra *lei* soa invariavelmente numa só silaba. Esta constatação significa que o hiato /u-a/ da primeira palavra, bem como o ditongo /ey/ da segunda, são as únicas formas possíveis de pronúncia na articulação dos referidos encontros em tais palavras, tratando-se por isso de encontros estáveis. Pelo contrário, podemos referir a palavra *luar* como exemplo de instabilidade. Note-se que durante uma pronúncia normal, essa mesma palavra, comporta-se como dissílaba, mas quando emitida rapidamente transforma-se em monossílaba, pela transformação do hiato /u-a/ no ditongo /wa/.

❍ **Encontros Consonânticos**

Tal como com as vogais, a língua portuguesa também permite encontros de consoantes — no máximo três.

Não se deve, no entanto, confundir consoantes e vogais com letras, uma vez que estas são apenas sinais que representam os referidos sons. Dessa forma os agrupamentos *bl* (de ***bl****oco*), *cr* (de ***cr****avo*), *pt* (de *a****pt****o*) e *gn* (de *di****gn****o*), por exemplo, representam encontros consonânticos. Já nas palavras *ca****rr****o*, *pê****ss****ego*, ***ch****ave*, *ma****lh****o* e *ca****nh****oto* não há qualquer encontro consonântico, pois os agrupamentos *rr*, *ss*, *ch*, *lh*, e *nh* representam uma só consoante. O mesmo acontece com as palavras *campo* e *ponto*, onde o *m* e o *n* são apenas sinal de nasalidade da vogal anterior, tal como se depreende das respectivas pronúncias, /kãpu/ e /põtu/.

Devem-se ainda distinguir dois tipos de encontros consonânticos:

- aqueles onde a primeira das consoantes termina sílaba — /r/, /l/ ou /x/ —, e por isso é seguida de outras sem constituírem, no entanto, um grupo (exs. /rt/ de *po****rt****as*, /rj/ de *fo****rj****a*, /rx/ de *ma****rch****a*, /ls/ de *sa****ls****a*, /lrr/ de *me****lr****o*, /xt/ de *fe****st****a* e /xk/ de *mú****sc****ulo*);

- aqueles onde duas consoantes pertencem à mesma sílaba e dessa forma constituem um "grupo próprio". Este é constituído pela fricativa /f/ ou por uma qualquer oclusiva, seguida por /l/ ou /r/ (exs. /tr/ de *ex****tr****a*, /br/ de *a****br****a*, /kl/ de *te****cl****a*, /pl/ de ***pl****ural*, /fr/ de ***fr****io*, e /fl/ de ***fl****orir*), e pode seguir-se a uma consoante final de sílaba (exs. /ltr/ de *u****ltr****a*, /rpl/ de *pe****rpl****exo* e /xkr/ de *de****scr****ever*).

Numa palavra não pronunciamos os fonemas separadamente; no entanto, quando emitida lentamente, dividimo-la numa sequência de pequenos grupos, que serão tantos quantas forem as vogais. Cada um desses grupos de sons representa uma sílaba, podendo ser formada por uma vogal ou um ditongo, acompanhados ou não de consoantes.

Numa palavra conseguimos normalmente distinguir uma sílaba acentuada — sílaba tónica — das restantes não acentuadas — sílabas átonas. A percepção distinta destes dois tipos de sílabas provém da maior ou menor dosagem de certas qualidades físicas:

- da INTENSIDADE, que representa a força expiratória com que os sons são pronunciados – resultando sons fortes (tónicos) ou fracos (átonos);
- do TOM, que representa a frequência de vibração das cordas vocais – resultando sons agudos (altos) ou graves (baixos);
- do TIMBRE, que representa o conjunto sonoro do tom fundamental com os seus armónicos produzidos pela ressonância daquele nas cavidades por onde passa o ar (tracto vocal e eventualmente nasal) – resultando sons abertos ou fechados;
- da QUANTIDADE, que traduz a duração com que os sons são emitidos – resultando sons longos ou breves.

Capítulo 3

Produção de Fala

As técnicas utilizadas nos codificadores de fala convencionais exploram, por norma, os princípios acústicos relacionados com o mecanismo humano de produção de fala, de modo a extraírem apenas aquilo que é essencial num sinal de voz para que possa ser reconstruído com alguma fidelidade a partir da menor quantidade de informação possível. Por conseguinte, a implementação ou estudo de um codificador de voz requer algum conhecimento prévio sobre o processo natural de produção de fala.

3.1 O Aparelho Fonador

A produção de fala depende dos órgãos e do sistema de respiração, pois os sons emitidos resultam quase sempre da acção desses órgãos sobre a corrente de ar vinda dos pulmões. É o ar que expiramos — a inspiração está normalmente relacionada com instantes de silêncio — que, quando sujeito a variações de pressão e volume pelos órgãos do sistema de produção de fala, cria as ondas sonoras que caracterizam a fala.

São essencialmente três as funções desempenhadas pelos órgãos do sistema humano durante o processo de produção de fala:

- garantirem o fluxo de ar;
- comportarem-se como obstáculos à passagem do ar; e
- formarem uma caixa de ressonância.

São estas três condições que em conjunto dão origem às ondas sonoras típicas da fala.

Na Figura 3-1 encontra-se representado o aparelho fonador, incluindo os seus principais órgãos, que podem ser agrupados do seguinte modo:

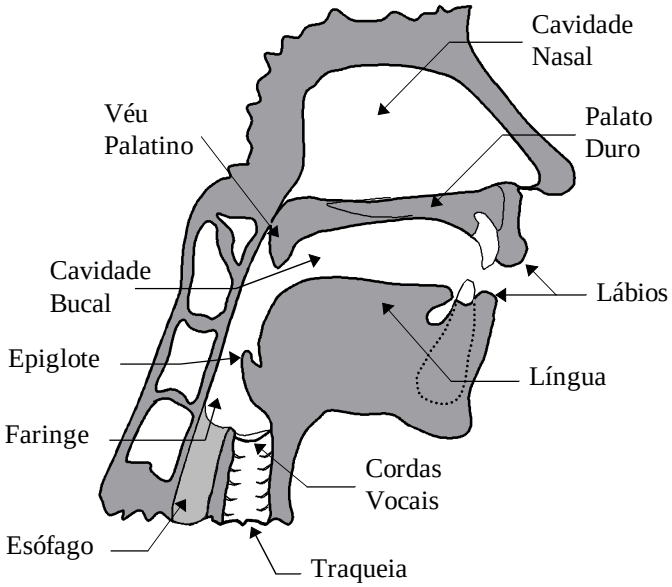


Figura 3-1: Secção transversal do aparelho fonador.

♦ *PULMÕES, BRÔNQUIOS e TRAQUEIA* — são os órgãos respiratórios que fornecem a corrente de ar necessária à fonação;

♦ *LARINGE* — zona onde se localizam as cordas vocais, responsáveis pelo tipo de excitação fornecido às cavidades superiores;

♦ *FARINGE, BOCA e FOSSAS NASAIS* — cavidades supralaríngeas que funcionam como caixas de ressonância.

O percurso da voz é idêntico ao da respiração. Durante a inspiração o volume dos pulmões aumenta, provocando dessa forma a diminuição da pressão do ar e a consequente entrada do mesmo para os pulmões. No fim da inspiração, a pressão do ar no interior dos pulmões volta a ser igual à pressão atmosférica, e o fluxo de ar pára. De seguida o processo inverte-se: o volume dos pulmões diminui e a pressão do ar aumenta em relação à pressão atmosférica, dando origem ao fluxo de ar, dos pulmões para o exterior. Nesta fase, o ar enviado pelos pulmões segue através da traqueia e laringe em direcção à faringe, que dá ligação às cavidades nasal e bucal. Os órgãos de fonação, dispostos ao longo do percurso do fluxo do ar, introduzem durante a fase de expiração uma maior pressão no ar circulante, tornando-o dessa forma audível.

A laringe é, sem dúvida, um dos órgãos mais importantes na fonação. É constituída por cartilagens, e encontra-se suspensa por membranas e músculos, podendo-se mover para cima e para baixo e assim alterar ligeiramente o volume das cavidades supraglotais e consequentemente a pressão do ar nessas cavidades. No interior da laringe encontram-se as cordas vocais, formadas por duas pregas

musculares localizadas nas paredes superiores da laringe. A glote representa a abertura entre essas duas pregas, que constituem o primeiro obstáculo ao fluxo de ar.

Durante a respiração normal, a glote encontra-se sempre aberta — os bordos das cordas vocais encontram-se separados —, permitindo dessa forma a livre circulação do ar. Durante a fonação, as cordas vocais vibram, abrindo e fechando rapidamente a passagem ao fluxo de ar vindo dos pulmões. A junção das cordas vocais — glote fechada — conduz ao aumento da pressão subglotal até a um ponto que obriga ao afastamento das cordas, uma da outra. Ao se afastarem, as cordas permitem a libertação de ar e a consequente diminuição da pressão subglotal, voltando dessa forma a aproximarem-se. A repetição consecutiva deste ciclo traduz a vibração das cordas vocais, durante a fonação.

3.2 O Tracto Vocal

O ar emitido pelos pulmões, em virtude do seu movimento, possui energia cinética, que se transforma em energia potencial caso o seu movimento seja impossibilitado por uma qualquer obstrução momentânea ao nível da cavidade bucal. Não sendo audível nenhuma dessas formas de energia, os órgãos de fonação têm como função convertê-las em energia acústica na forma de ondas sonoras. Essa conversão é conseguida por interposição de alguns órgãos, vibratórios ou não, no trajecto da corrente expiratória, ou então, através de um qualquer estreitamento ao nível do tracto vocal.

Uma vez que a produção do som vocal depende das interferências na corrente de ar impostas ao longo do tracto vocal — mais concretamente desde a glote até aos lábios, pois é entre estes extremos que se realiza a conversão de energia cinética em acústica —, este durante a produção do som deve comportar-se, de alguma forma, como um tubo de ressonância.

São utilizados, basicamente, dois tipos de modelos para estudo e representação das ressonâncias do tracto vocal — cavidades faríngea e bucal. O primeiro consiste precisamente em representar as cavidades do tracto vocal pela concatenação de uma série de tubos ressonantes de distintas secções, e confrontar o efeito da variação do volume ou da área das aberturas de cada um destes tubos com as ressonâncias globais resultantes. Sabe-se que o aumento do volume dum tubo simples, tem como

consequência, a diminuição da sua frequência de ressonância, e o aumento da área das aberturas, eleva-a. Contudo, numa junção de dois ou mais tubos, cada um afecta o modo de vibração do ar em todos os outros, e por isso, esse efeito de acoplamento faz com que a interpretação das frequências de ressonância, à custa do efeito da variação das características físicas dos tubos individuais, se torne relativamente complexa.

O segundo modelo consiste em representar o tracto vocal por uma linha de transmissão ressonante. Esta, como se sabe, consiste num sistema que transmite energia de um ponto para outro na forma de uma onda progressiva com um comprimento de onda inferior ou, pelo menos, comparável ao comprimento físico do sistema. Embora este tipo de tratamento das ressonâncias do tracto vocal pareça mais correcto, é no entanto muito mais complexo, não permitindo estabelecer quaisquer conclusões simples na relação entre as configurações do tracto vocal e as frequências obtidas nas ondas acústicas.

3.2.1 O Tubo Acústico

Todo o som, mesmo o produzido vocalmente, quando propagado no ar exterior já não é o mesmo que foi produzido na fonte; antes, sofreu alterações impostas pela ressonância do tracto vocal. Assim, as cavidades superiores do aparelho fonador comportam-se como um tubo acústico responsável pela modulação das ondas sonoras provenientes da laringe. Este tubo tem início na abertura posterior da laringe, ou glote, e termina nos lábios. A área da sua secção transversal, determinada pelas posições dos seus elementos articuladores — lábios, bochechas, língua, maxilar inferior e véu palatino —, pode variar entre zero e cerca de 20 cm^2.

Qualquer tubo percorrido por um fluxo de ar, se convenientemente excitado, actua como um ressoador, e como tal, a sua resposta a cada diferente frequência dependerá fundamentalmente do seu volume interior e da sua forma. Assim, como as características de ressonância de um tubo acústico dependem das suas características físicas — volume, forma, etc. —, a variação destas resulta na modificação do tipo de ressonância imposta pelo tubo no ar que por ele circula. Tal variação no tracto vocal é função do movimento muscular, mais concretamente, da actividade dos seus órgãos articulatórios.

3.2.2 Ressonâncias do Tracto Vocal (formantes)

A diferença entre o espectro das ondas acústicas propagadas no ar exterior e o espectro das ondas emitidas ao nível da glote, resultam das características de transmissão específicas do tracto vocal. Essas características assemelham-se a um filtro com uma complexa resposta na frequência, tal como a que se ilustra na Figura 3-2b. Dessa forma, ondas acústicas dentro de determinadas gamas de frequências são transmitidas, enquanto que outras, fora dessas gamas, são essencialmente atenuadas. Assim a resposta na frequência do tracto vocal é caracterizada por uma série de regiões, que passaremos a designar por formantes, responsáveis pela transmissão das ondas acústicas sem atenuação relevante.

As formantes encontram-se relacionadas com os picos visualizados no espectro de um sinal de voz, e são normalmente identificadas de acordo com as suas posições relativas ao longo do eixo da frequência: formante 1, formante 2, etc. Assim, se na Figura 3-2 o espectro em a) pertencer ao sinal de excitação e o filtro em b) representar a resposta na frequência do tracto vocal, então o espectro do sinal acústico resultante será como o dado em c). Repare-se na existência de três formantes neste exemplo ilustrativo: formantes 1, 2 e 3, centradas respectivamente a cerca de 500 Hz, 1200 Hz e 2000 Hz. Para a maioria das configurações do tracto vocal existem três a cinco formantes inferiores a 5 KHz. As amplitudes e localizações das três primeiras, que ocorrem normalmente abaixo dos 3 KHz, são extremamente importantes na síntese e percepção da fala. Por sua vez, as formantes superiores são importantes essencialmente para a representação da fala não vozeada.

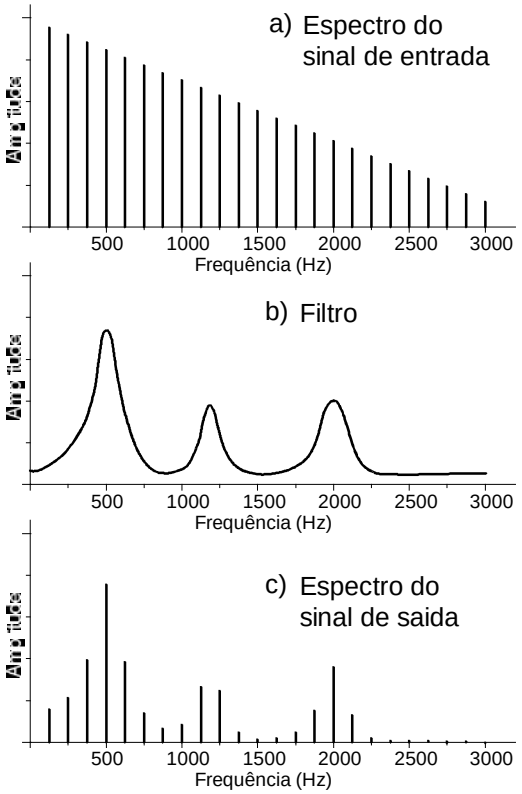


Figura 3-2: Processo ilustrativo da filtragem por parte do tracto vocal.

Embora o efeito ressonante do tracto vocal se faça sentir na produção de todos os sons, e por isso as formantes estejam sempre presentes, é por vezes difícil em alguns tipos de sons a sua distinção e localização na frequência. Pois, pode acontecer que a fonte de potência acústica de um som particular não forneça energia nas frequências situadas nas bandas de passagem da função de transferência do tracto vocal, ou então, a potência de excitação nessas bandas ser consideravelmente inferior à potência nas outras frequências, e dessa forma as potências das frequências formantes do som produzido não ser suficiente de modo a se realçar das restantes.

Através do posicionamento e da energia das formantes, podemos identificar certas características físicas dos sons. Os sons graves, por exemplo, caracterizam-se por uma maior concentração de energia na parte inferior do espectro. Contrariamente, nos sons agudos a energia predomina nas componentes da parte superior do espectro. Constata-se que as zonas de maior concentração de energia resultam, na maior parte dos casos, da proximidade de duas ou mais formantes. Se a segunda formante se encontrar próxima da primeira, o som é, normalmente, grave. Se pelo contrário, a mesma se encontrar junto à terceira formante, estaremos muito provavelmente na presença de um som agudo. A proximidade da terceira e quarta formantes pode ainda ser suficiente para determinar o carácter agudo do som.

Sabendo que, a cada configuração do tracto vocal corresponde, no sinal de voz, um grupo de frequências de ressonância designadas por formantes, e como a configuração do tracto vocal varia com o tempo, as propriedades espectrais do sinal de voz também variam, tratando-se por isso de um sinal não estacionário.

3.3 Acoplamento Nasal

Na parte superior da faringe (Figura 3-1) existe uma encruzilhada que oferece ao fluxo de ar duas vias alternativas de acesso ao exterior: o canal bucal e o nasal. Existe também uma membrana dotada de alguma mobilidade, designada por véu palatino, capaz de obstruir o acesso ao canal nasal. Desse modo, quando o véu palatino se encontra colado à parede superior da faringe, os sons articulados denominam-se orais, visto ser o tracto vocal — faringe e cavidade bucal — a única cavidade de ressonância utilizada na produção do som. Quando se encontra deslocado para baixo, o véu palatino deixa ambas as passagens livres, e o tracto vocal é acoplado acusticamente

com a cavidade nasal que funciona como um outro canal de passagem, introdutor de uma ressonância típica na vibração do ar que por lá passa. Produzindo-se dessa forma os sons nasais.

Ao nível espectral, a utilização do canal nasal como cavidade de ressonância adicional, tem como consequência a introdução de anti-ressonâncias e de ressonâncias adicionais, que provocam alterações nas formantes relacionadas com a configuração do tracto vocal, durante a produção dos sons nasalados. Estas alterações manifestam-se sobretudo num aumento da largura de banda e na diminuição da energia das formantes, em especial da primeira (formante 1). Um espectro com anti-ressonâncias é ainda caracterizado por potências nulas, ou pelo menos insignificantes em determinadas frequências.

3.4 Tipos de Excitação

A fala é composta por ondas acústicas provenientes do sistema fonador quando o fluxo de ar expelido pelos pulmões é modulado pelas cavidades de ressonância — tracto vocal e eventualmente nasal — e sofre perturbações pelos órgãos articuladores do sistema. Os pulmões são, portanto, os responsáveis pelo fornecimento da potência de excitação no sistema acústico humano. Neste sistema existem essencialmente três mecanismos responsáveis pela produção de ondas acústicas:

♦ um processo através do qual o ar é forçado a atravessar a glote com a tensão das cordas vocais ajustada de modo a vibrarem, produzindo-se dessa forma impulsos de ar quase periódicos que vão excitar o tracto vocal;

♦ formação de uma constrição ao nível da cavidade bucal, de modo a que, forçando o ar a passar por esse estreitamento com uma certa velocidade, se produza turbulência;

♦ impor uma obstrução completa ao nível da cavidade bucal de forma a gerar-se uma elevada pressão atrás dessa obstrução que ao ser anulada abruptamente, resulte uma excitação transitória.

Consoante o tipo de excitação, podem-se obter sons vozeados, ou não vozeados. Se houver vibração das cordas vocais, o som é vozeado, caso contrário o som resulta não vozeado. Os sons vozeados correspondem a sinais no domínio do tempo

aproximadamente periódicos. Ao período correspondente designa-se por *pitch*, e a respectiva frequência por frequência fundamental. No domínio da frequência, os sons vozeados são caracterizados por linhas espectrais ou harmónicas e apresentam normalmente uma estrutura de formantes nítida. São caracterizados ainda por uma forte componente de baixa frequência que se manifesta nos espectrogramas sob a forma de uma barra horizontal próxima do eixo das abcissas. Os sons não vozeados, pelo contrário, correspondem a sinais, no domínio do tempo, de natureza aproximadamente aleatória e com espectros de elevada largura de banda. Adicionalmente, constata-se que as zonas não vozeadas de um sinal de voz têm em geral menos energia do que as zonas vozeadas.

Em suma, podemos afirmar que a estrutura harmónica de um sinal de voz é uma consequência da quase periodicidade do sinal e pode ser atribuída à vibração das cordas vocais. A envolvente espectral (estrutura formante) deve-se à interacção entre a fonte de excitação e o tubo acústico formado pelo tracto vocal e, eventualmente, nasal.

Capítulo 4

Modelo Digital de Produção de Voz

Anteriormente abordaram-se alguns aspectos da teoria acústica relacionados com a produção de fala, pois a sua compreensão torna-se fundamental sempre que se pretenda aplicar técnicas de processamento digital em sinais de voz. No presente capítulo veremos como esses conceitos acústicos nos podem conduzir a modelos digitais de representação de sinais de voz amostrados. Esses modelos serão obtidos usando o método clássico que modela um sinal voz como sendo o resultado da saída de um sistema linear variável no tempo, excitado quer por ruído aleatório, quer por uma sequência de impulsos aproximadamente periódica.

4.1 Modelo de Excitação

Um modelo detalhado da excitação do sistema vocal envolveria necessariamente equações diferenciais demasiado complexas. Contudo, servindo-nos apenas dos princípios básicos da geração do som, obtemos um modelo, embora simples, largamente utilizado na produção de voz sintética.

Como já referido anteriormente, as ondas acústicas emitidas pelo sistema vocal são geradas essencialmente por três mecanismos distintos: vibração das cordas vocais; turbulência provocada por uma zona de constrição; e oclusão momentânea da cavidade bucal.

♦ Como sabemos, a vibração das cordas vocais confere a excitação para os sons vozeados. A frequência de oscilação — taxa a que a glote se fecha e se abre — é controlada pela pressão do ar nos pulmões, pela tensão e elasticidade das cordas vocais, e pela área da abertura glotal em condições de repouso. A estas ainda se deve acrescentar a influência do tracto vocal, uma vez que as variações de pressão no interior do tracto vocal influenciam as variações de pressão na glote. Em termos de

análogo eléctrico, o tracto vocal comporta-se como uma carga sobre o oscilador glotal, tal como ilustrado na Figura 4-1.

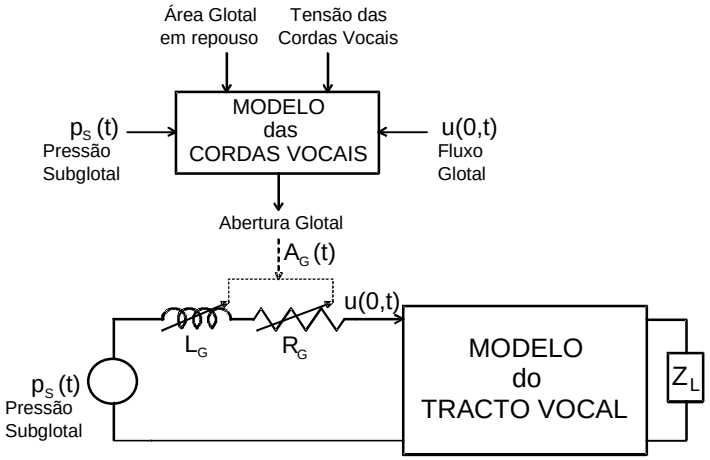


Figura 4-1: Modelo do sistema vocal para sons vozeados.

O acoplamento entre o modelo representativo das cordas vocais e o modelo referente ao tracto vocal é representado por intermédio de uma resistência e de uma indutância acústica variáveis, controladas pela função $1/A_G(t)$, onde $A_G(t)$ expressa a evolução no tempo da abertura glotal. Note-se, por exemplo, que quando a glote se encontra totalmente fechada ($A_G(t)=0$) a impedância é infinita, resultando por isso um fluxo de ar nulo à entrada do tracto vocal. Constata-se que o fluxo glotal é ciclicamente interrompido, dando origem a impulsos de ar quase periódicos. Na Figura 4-2 encontra-se ilustrado um exemplo deste tipo de sinal.

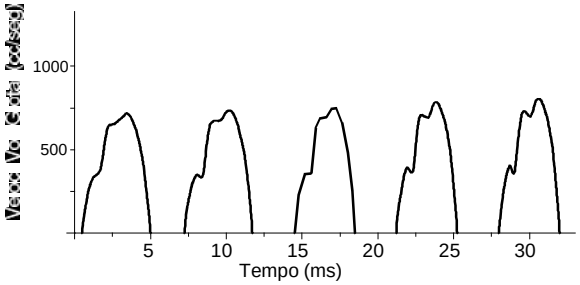


Figura 4-2: Ilustração do fluxo glotal referente a um som vozeado.

Encarando o sistema tal como representado na Figura 4-1, o sistema vocal seria necessariamente não linear. Contudo, uma vez que o acoplamento entre a glote e o tracto vocal é fraco, pode ser desprezado, permitindo assim a separação da excitação,

do sistema de transmissão (tracto vocal) e consequentemente a linearização das partes, resultando o modelo simplificado expresso na Figura 4-3.

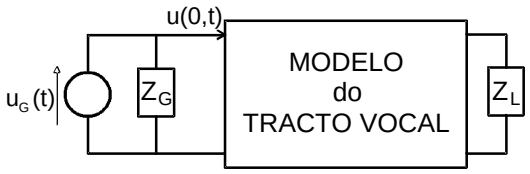


Figura 4-3: Modelo simplificado do sistema vocal.

Neste caso, $u_G(t)$ corresponde a uma "fonte de fluxo" cuja forma de onda poderá ser, por exemplo, a que se encontra na Figura 4-2; e Z_G é a impedância acústica dada por $Z_G(w) = R_G + jwL_G$, com R_G e L_G constantes. A interligação entre os modelos de excitação e tracto vocal passou a ser representada no domínio da frequência pela seguinte relação

$$U(0,w) = U_G(w) - P(0,w)/Z_G(w), \qquad (4\text{-}1)$$

com $P(0,w)$ representando a transformada de Fourier da pressão acústica à entrada do tracto vocal.

Verifica-se que a impedância $Z_G(w)$ aumenta com a frequência [Rabiner (79)]. Por isso, a impedância glotal tem como efeito aumentar a largura de banda das ressonâncias mais baixas do sistema de produção de voz.

♦ Os sons fricativos não vozeados resultantes da turbulência do fluxo de ar ao passar por entre um estreitamento, merecem um tratamento diferenciado. Neste caso, como as cordas vocais não vibram, a excitação pode ser modelada por uma fonte de ruído aleatório, com uma potência intimamente relacionada com a intensidade do fluxo de ar no tubo. Para além da turbulência provocada pela constrição, as cordas vocais poderão simultaneamente vibrar, resultando dessa forma um som misto. É o caso das consoantes fricativas vozeadas que, devido à sua natureza, terão um modelo de excitação misto, combinando uma fonte de ruído com uma fonte de impulsos semelhantes aos da Figura 4-2.

♦ Finalmente deveria-se ainda considerar os sons plosivos ou oclusivos que resultam da obstrução completa da cavidade bucal — quando liberta abruptamente, resulta numa excitação transitoriamente turbulenta —, bem como os sons de natureza

vibrante. Para estes não é possível estabelecer um modelo de excitação aceitável, pois sendo sons descontínuos, as suas características espectrais alteram-se drasticamente ao longo do tempo.

4.2 Modelação do Tracto Vocal

O som, tratando-se de uma vibração, propaga-se no ar ou através de qualquer outro meio, pela própria vibração das suas partículas. Assim, a geração e consequente propagação no tracto vocal de ondas acústicas obedecem necessariamente a leis físicas, nomeadamente a leis de conservação de massa, conservação de energia, de termodinâmica e fluidos. Uma vez que o ar é o principal meio de propagação do som no sistema vocal, aplicando os princípios físicos associados a um fluido de fraca viscosidade é possível obter-se um conjunto de equações diferenciais parciais que, de alguma forma, descrevam a progressão do ar no sistema vocal [Portnoff (73)] [Sondhi (74)]. Porém, a formulação e resolução dessas equações é extremamente complexa, a não ser que se utilizem simplificações significativas sobre a forma do tracto vocal e perdas de energia em todo o sistema vocal.

Uma formulação detalhada levaria em conta todos os seguintes aspectos:

- Variação da forma do tracto vocal, no tempo;
- Perdas devido à condutividade térmica e fricção de viscosidade nas paredes do tracto vocal;
- Elasticidade das paredes do tracto vocal;
- Radiação do som pelos lábios;
- Acoplamento nasal;
- Excitação do tracto vocal.

De uma forma simplista, o tracto vocal pode ser representado por um tubo de secção transversal não uniforme e variável no tempo. Para sons com frequências correspondentes a comprimentos de onda consideravelmente elevados quando comparados com as próprias dimensões do tracto vocal — é o caso de frequências inferiores a cerca de 4 KHz —, podemos também admitir existir apenas ondas planas

a propagarem-se ao longo do tubo vocal. Desprezando igualmente as perdas de viscosidade e de condutividade térmica, é possível demonstrar [Portnoff (73)] que as ondas acústicas no interior de um tubo com as características referidas satisfazem as seguintes equações diferenciais:

$$-\frac{\partial p(x,t)}{\partial x} = \rho \frac{\partial (u(x,t)/A(x,t))}{\partial t}, \tag{4-2a}$$

$$-\frac{\partial u(x,t)}{\partial x} = \frac{1}{\rho c^2} \frac{\partial (p(x,t)A(x,t))}{\partial t} + \frac{\partial A(x,t)}{\partial t}, \tag{4-2b}$$

onde $p(x,t)$ representa a pressão acústica no interior do tubo na posição x e instante t; $u(x,t)$ o fluxo em função da posição x e do instante t; $A(x,t)$ a área da secção transversal do tubo como função da distância x e do instante t; ρ a densidade do ar no interior do tubo e c a velocidade do som.

A resolução das equações diferenciais (4-2) subentende a determinação da pressão $p(x,t)$ e do fluxo $u(x,t)$ ao longo do tempo em toda a extensão do tracto vocal. Para isso á necessário que sejam impostas condições de fronteira referentes a cada um dos extremos do tracto vocal — respectivamente, glote e lábios —, dependentes quer da natureza da excitação ao nível da glote, quer do efeito de radiação do som pelos lábios. Obviamente que a área transversal do tracto vocal $A(x,t)$ terá de ser igualmente conhecida. Pretensão algo difícil de conseguir, uma vez que, mesmo para sons contínuos — situação durante a qual o tracto vocal não sofre alterações significativas na sua forma —, torna-se extremamente difícil fazer uma medição precisa de $A(x,t)$. Uma das técnicas utilizadas para esse efeito baseia-se na análise de imagens por Raio-X. A forma do tracto vocal tem sido também estimada a partir de medidas acústicas obtidas através da excitação do tracto vocal por uma fonte externa. Contudo, nenhum destes métodos tem conduzido a medidas suficientemente precisas para poderem ser utilizados na sintetização de sinais de voz. Servem apenas para nos darem uma ideia sobre as grandezas envolvidas.

Infelizmente, mesmo com todos os parâmetros atrás mencionados completamente determinados, a resolução das equações (4-2) manter-se-ia extremamente complexa se outras restrições ou simplificações não fossem consideradas.

4.2.1 Tubo Acústico Uniforme Sem Perdas

Felizmente, constata-se que, a aceitação de um modelo extremamente simples, onde a função de área do tracto vocal, $A(x,t)$, seja constante ao longo de x e de t — tubo de secção transversal uniforme e invariável no tempo —, é suficiente para reproduzir as características típicas do tracto vocal. Embora este modelo seja demasiado simplista, é pertinente a sua consideração na medida em que o método de análise utilizado, bem como as características essenciais das soluções resultantes, têm muito em comum com modelos mais realistas, e além disso, um modelo bem mais realista pode ser obtido pela concatenação de vários desses tubos de secção uniforme.

Portanto, para um tubo de secção transversal uniforme e invariável no tempo as equações (4-2) simplificam-se nas seguintes

$$-\frac{\partial p(x,t)}{\partial x} = \frac{\rho}{A}\frac{\partial u(x,t)}{\partial t} \tag{4-3a}$$

$$-\frac{\partial u(x,t)}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p(x,t)}{\partial t}, \tag{4-3b}$$

onde a função de área passou a ser constante, $A(x,t) = A$; e como demonstrado em [Rabiner (79)], as soluções das equações (4-3) terão a seguinte forma

$$u(x,t) = u^{+}(t - x/c) - u^{-}(t + x/c), \tag{4-4a}$$

$$p(x,t) = \frac{\rho c}{A}\left[u^{+}(t - x/c) + u^{-}(t + x/c)\right]. \tag{4-4b}$$

Nestas equações, as funções $u^{+}(t - x/c)$ e $u^{-}(t + x/c)$ representam ondas progressivas com direcções de progressão opostas. Dependem das condições de fronteira, e por isso, da natureza da excitação ao nível da glote e do efeito de radiação pelos lábios.

De modo a estudarmos o comportamento do tubo uniforme sem perdas no domínio da frequência, precisamos de garantir a verificação de uma condição de fronteira na posição $x = 0$. Por exemplo, podemos admitir que o tubo é excitado com um fluxo $u(0,t)$ variável segundo uma exponencial complexa de frequência angular w e com uma amplitude complexa $\mathrm{U}(0,w)$, isto é

$$u(0,t) = \mathrm{U}(0,w)e^{jwt}. \tag{4-5}$$

Uma vez que as equações (4-3) são lineares, as funções $u^+(t - x/c)$ e $u^-(t + x/c)$ terão necessariamente a forma

$$u^+(t - x/c) = \mathrm{K}^+ e^{jw(t-x/c)}, \tag{4-6a}$$

$$u^-(t + x/c) = \mathrm{K}^- e^{jw(t+x/c)}. \tag{4-6b}$$

Não sendo conhecidas as constantes K^+ e K^-, existe ainda a necessidade de impor uma segunda condição de fronteira. Podemos admitir, por simplicidade, que a pressão acústica é nula na extremidade final do tubo, ou seja

$$p(l,t) = 0, \tag{4-7}$$

onde l representa o comprimento do tubo acústico. Por fim, substituindo as funções (4-6) nas equações (4-4), e utilizando as condições de fronteira (4-5) e (4-7) referentes às extremidades do tubo associadas às aberturas, respectivamente da glote e dos lábios, obtêm-se as seguintes soluções para $u(x,t)$ e $p(x,t)$:

$$u(x,t) = \frac{\cos\left[w(l-x)/c\right]}{\cos\left[wl/c\right]} \mathrm{U}(0,w)\, e^{jwt}, \tag{4-8a}$$

$$p(x,t) = j\frac{\rho c}{A}\frac{\sin\left[w(l-x)/c\right]}{\cos\left[wl/c\right]} \mathrm{U}(0,w)\, e^{jwt}. \tag{4-8b}$$

Portanto, estas equações expressam o fluxo $u(x,t)$ e a pressão $p(x,t)$ em qualquer ponto do tubo em função do fluxo sinusoidal $u(0,t)$ à entrada do tubo. Mas como estamos apenas interessados na relação existente entre o fluxo à entrada (glote) e o fluxo à saída (lábios) do tubo, isto é, entre $u(0,t)$ e $u(l,t)$, impondo $x = l$ obtem-se

$$u(l,t) = \frac{1}{\cos\left[wl/c\right]} \mathrm{U}(0,w)\, e^{jwt} = \mathrm{U}(l,w)\, e^{jwt}, \tag{4-9}$$

de onde resulta a seguinte relação de amplitudes

$$\mathrm{V}_a(jw) = \frac{\mathrm{U}(l,w)}{\mathrm{U}(0,w)} = \frac{1}{\cos\left[wl/c\right]} \tag{4-10}$$

que traduz precisamente a resposta na frequência do tubo acústico uniforme. Como exemplo, se considerarmos um tubo uniforme de 17,5 cm de comprimento e admitindo que a velocidade do som é 35000 cm/s, obtem-se a resposta na frequência ilustrada na Figura 4-4.

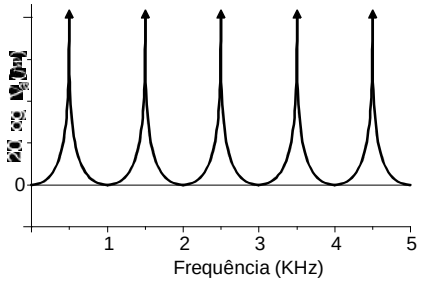


Figura 4-4: Exemplo de resposta na frequência referente a um tubo uniforme sem perdas.

De modo a obtermos uma representação da função $V_a(jw)$ em termos de transformadas de Laplace, substituímos a variável *w* por s/j. Facilmente se verifica que a função de transferência possui um número infinito de pólos igualmente espaçados ao longo do eixo *jw*. Concretamente

$$s_n = \pm j\frac{c}{l}(n+1/2)\pi, \quad n = 0,1,2,... \tag{4-11}$$

que, tal como ilustrado na Figura 4-4, correspondem às frequências de ressonância do sistema. São estas frequências que representam as formantes quando o sistema em causa está associado com a produção de fala.

Apoiando-nos na teoria de Fourier, fica claro que a relação encontrada na equação (4-10) não determina apenas o comportamento do sistema unicamente para sinusóides; antes, expressa a relação entre a transformada de Fourier do sinal de saída e a transformada de Fourier do sinal de entrada (resposta na frequência), traduzindo por isso o comportamento do sistema para entradas arbitrárias. Dessa forma, a resposta na frequência dada em (4-10) representa uma primeira caracterização do modelo para o sistema vocal.

4.2.2 Efeito da Radiação Labial

Na dedução da resposta na frequência referente a um único tubo sem perdas consideramos nula a pressão acústica à saída do tubo, isto é, impusemos a condição fronteira $p(l,t) = 0$. No análogo eléctrico esta situação corresponde a um curto-circuito. Na realidade trata-se de uma simplificação algo irrealista uma vez que o tracto vocal termina com uma abertura entre os lábios onde, para haver alterações no fluxo terá que ocorrer necessariamente variações equivalentes na pressão acústica.

Subentende-se por isso a existência duma impedância á saída do sistema acústico que fundamente a existência da relação entre a pressão e o fluxo na abertura labial, isto é

$$\mathrm{P}(l,w) = \mathrm{Z}_L(w)\mathrm{U}(l,w), \tag{4-12}$$

onde, como referido em [Rabiner (79)] a impedância de radiação, sendo dependente da frequência, pode ser aproximada por

$$\mathrm{Z}_L(w) = \frac{jwL_L R_L}{R_L + jwL_L}, \tag{4-13}$$

que traduz o paralelo duma resistência R_L com uma indutância de radiação L_L. Os valores

$$R_L = 128/(a\pi^2), \text{ e} \tag{4-14}$$

$$L_L = 8a/(3\pi c) \tag{4-15}$$

são considerados boas aproximações para as referidas grandezas (o *a* representa o raio da abertura entre os lábios e *c* a velocidade do som).

Pela análise da expressão (4-13) retira-se que para frequências bastante baixas $\mathrm{Z}_L(w) \approx 0$, o que significa ser esta a situação em que a impedância de radiação se comporta aproximadamente como um curto-circuito ($p(l,t) = 0$). Para altas frequências, em que $w >> R_L/L_L$, a impedância é puramente real podendo ser aproximada por $\mathrm{Z}_L(w) \approx R_L$. Finalmente, para uma gama intermédia de frequências, embora suficientemente baixas de modo a que $w << R_L/L_L$, a indutância é essencialmente imaginária, podendo ser aproximada por $\mathrm{Z}_L(w) \approx jwL_L$.

Decompondo a impedância dada na equação (4-13) nas suas componentes, real e imaginária,

$$\mathrm{Z}_L(w) = \frac{L_L^2 R_L}{R_L^2/w^2 + L_L^2} + j\frac{wL_L R_L^2}{R_L^2 + w^2 L_L^2}, \tag{4-16}$$

verifica-se que a sua parte real aumenta com a frequência. Assim, uma vez que a energia dissipada é proporcional apenas à componente real da impedância, podemos concluir que no modelo completo de produção de voz as perdas de radiação aumentam com a frequência.

Note-se também que a introdução duma impedância de radiação à saída do tubo influencia o tipo de onda que o percorre. Nomeadamente, a carga de radiação tem como efeito aumentar a largura de banda das formantes — fundamentalmente nas

altas frequências — devido ao aumento das perdas, e baixar ligeiramente as frequências de ressonância (deslocamento das formantes).

4.2.3 Modelo dos Tubos Acústicos

Um modelo bastante utilizado para a produção de voz baseia-se na suposição de que o tracto vocal pode ser caracterizado por uma conjunto de tubos acústicos uniformes de diferentes secções acoplados em série, tal como ilustrado na Figura 4-5.

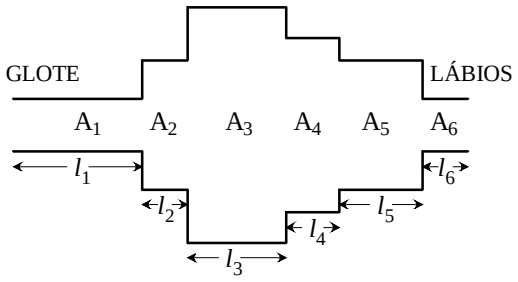


Figura 4-5: Ilustração da concatenação de 6 tubos acústicos.

As áreas da secção transversal dos tubos, representadas por $\{A_k\}$, são escolhidas de forma a se aproximarem da função de área efectiva, $A(x)$, do tracto vocal. Se na concatenação utilizarmos um elevado número de tubos de comprimento reduzido é de esperar que a resposta na frequência do modelo resultante esteja muito próxima da resposta de um tubo com a função de área contínua.

A partir das equações (4-4) facilmente identificamos as equações que descrevem a propagação do som em cada um dos tubos. Assim, para o *k* ésimo tubo

$$u_k(x,t) = u_k^+(t - x/c) - u_k^-(t + x/c), \tag{4-17a}$$

$$p_k(x,t) = \frac{\rho c}{A_k}\left[u_k^+(t - x/c) + u_k^-(t + x/c)\right], \tag{4-17b}$$

onde *x* representa a distância medida a partir do estremo inicial do *k* ésimo tubo. De modo a obtermos a relação existente entre as ondas progressivas em tubos adjacentes, necessitamos encontrar condições de fronteira referentes a ambos os extremos de cada um dos tubos. Para isso devemos recorrer aos princípios físicos de que tanto a pressão como o fluxo são funções contínuas ao longo do tempo e no espaço.

Considerando que $\tau_k = l_k / c$ representa o tempo que uma onda progressiva demora a percorrer o *k* ésimo tubo, a Figura 4-6 ilustra a junção entre o *k* ésimo e o (*k*+1) ésimo tubos.

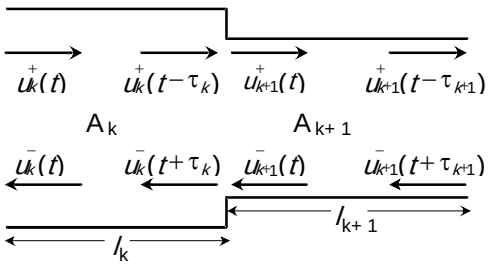


Figura 4-6: Junção entre dois tubos acústicos.

Apenas parte da onda progressiva positiva que atinge a junção, $u_k^+(t-\tau_k)$, é transmitida para o tubo seguinte, enquanto que a restante é reflectida para trás. De igual forma, parte da onda negativa que atinge a junção, $u_{k+1}^-(t)$, é transmitida, sendo a restante reflectida. Por outro lado, os princípios de continuidade resultam nas seguintes condições de fronteira

$$u_k(l_k,t) = u_{k+1}(0,t), \tag{4-18a}$$

$$p_k(l_k,t) = p_{k+1}(0,t). \tag{4-18b}$$

Substituindo as equações (4-17) nas equações (4-18) obtem-se

$$u_k^+(t-\tau_k) - u_k^-(t+\tau_k) = u_{k+1}^+(t) - u_{k+1}^-(t), \tag{4-19a}$$

$$\frac{A_{k+1}}{A_k}\left[u_k^+(t-\tau_k) + u_k^-(t+\tau_k)\right] = u_{k+1}^+(t) + u_{k+1}^-(t). \tag{4-19b}$$

Ficamos portanto em condições de expressar as ondas progressivas que deixam a junção em função das ondas que atingem a mesma, ou seja, $u_k^-(t+\tau_k)$ e $u_{k+1}^+(t)$ em função de $u_k^+(t-\tau_k)$ e de $u_{k+1}^-(t)$. Depois de resolvidas as equações (4-19), obtem-se

$$u_{k+1}^+(t) = \left(\frac{2A_{k+1}}{A_{k+1}+A_k}\right)u_k^+(t-\tau_k) - \left(\frac{A_k - A_{k+1}}{A_{k+1}+A_k}\right)u_{k+1}^-(t), \tag{4-20a}$$

$$u_k^-(t+\tau_k) = \left(\frac{A_k - A_{k+1}}{A_{k+1}+A_k}\right)u_k^+(t-\tau_k) + \left(\frac{2A_k}{A_{k+1}+A_k}\right)u_{k+1}^-(t), \tag{4-20b}$$

de onde se constata que a fracção da onda $u_k^+(t-\tau_k)$ que é reflectida na junção é dada por

$$r_k = \left(\frac{A_k - A_{k+1}}{A_{k+1} + A_k} \right) \tag{4-21}$$

e obviamente, a porção de onda $u_{k+1}^-(t)$ reflectida é $-r_k$. Portanto, r_k representa o coeficiente de reflexão da *k* ésima junção. Podemos assim expressar as equações (4-20) de forma mais compacta,

$$u_{k+1}^+(t) = (1-r_k)u_k^+(t-\tau_k) - r_k u_{k+1}^-(t), \tag{4-22a}$$

$$u_k^-(t+\tau_k) = r_k u_k^+(t-\tau_k) + (1+r_k)u_{k+1}^-(t). \tag{4-22b}$$

A partir destas equações facilmente convertemos o esquema da Figura 4-6 num diagrama de fluxo representando cada uma das junções do modelo da Figura 4-5. Como neste exemplo o modelo é constituido por 6 tubos, o mesmo será representado por 5 junções do tipo da que se encontra na Figura 4-7, cada uma das quais caracterizada por um coeficiente de reflexão.

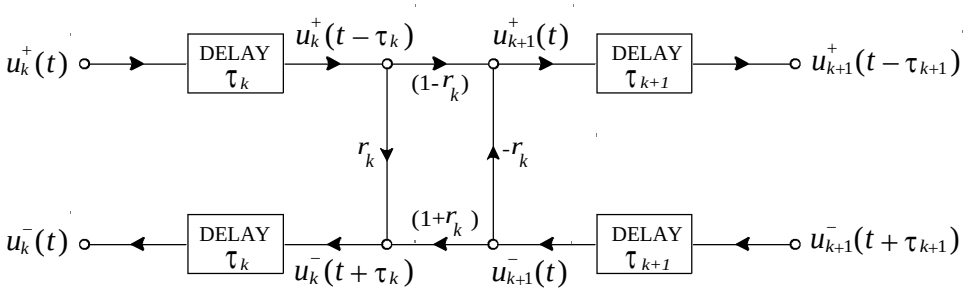


Figura 4-7: Diagrama de fluxo representando a junção entre dois tubos acústicos.

Desta forma é já possível relacionar as ondas progressivas da entrada com as de saída do modelo constituído por vários tubos acústicos. No entanto o nosso interesse recai, não na relação entre essas grandezas, mas sim entre os valores de pressão e fluxo à entrada e os correspondentes valores à saída do modelo. Por isso necessitamos de recorrer às condições de fronteira referentes aos extremos do sistema vocal para encontrarmos as expressões que relacionam as grandezas pretendidas com as respectivas ondas progressivas.

Assumindo um modelo constituido por *N* tubos, a condição de fronteira associada aos lábios relaciona a pressão radiada com o fluxo na extremidade do último tubo, isto é, relaciona $p_N(l_N,t)$ com $u_N(l_N,t)$. O tracto vocal termina numa

abertura entre os lábios, correspondendo no modelo eléctrico a uma impedância de radiação Z_L que se, por simplicidade, a considerarmos puramente real ($Z_L = R_L$), obtem-se a seguinte relação

$$p_N(l_N, t) = R_L u_N(l_N, t), \tag{4-23}$$

que, depois de substituídos os valores dados nas equações (4-17), vem

$$\frac{\rho c}{A_N}\left[u_N^+(t-\tau_N) + u_N^-(t+\tau_N)\right] = R_L\left[u_N^+(t-\tau_N) - u_N^-(t+\tau_N)\right], \tag{4-24}$$

resultando

$$u_N^-(t+\tau_N) = \frac{R_L - \rho c/A_N}{R_L + \rho c/A_N} u_N^+(t-\tau_N) = r_L u_N^+(t-\tau_N), \tag{4-25}$$

onde

$$r_L = \frac{R_L - \rho c/A_N}{R_L + \rho c/A_N} \tag{4-26}$$

passa a representar o coeficiente de reflexão associado à abertura labial. Portanto o fluxo á saída do sistema será dado por

$$u_N(l_N, t) = u_N^+(t-\tau_N) - u_N^-(t+\tau_N) = (1-r_L)u_N^+(t-\tau_N). \tag{4-27}$$

Note-se que, para simplificar, assumimos que a impedância de carga Z_L era real. Caso contrário, teria que se substituir a equação (4-25) pela sua equivalente no domínio da frequência, e daí resultaria necessariamente um coeficiente de reflexão complexo.

A condição de fronteira associada à fonte de excitação relaciona a pressão com o fluxo na extremidade inicial do primeiro tubo acústico, isto é, relaciona $p_1(0,t)$ com $u_1(0,t)$. Recorrendo ao modelo simplificado da Figura 4-3, onde a fonte de excitação é linearmente separável do tracto vocal, e admitindo novamente por simplicidade que a impedância Z_G é real e igual a R_G, encontramos a seguinte relação

$$u_1(0,t) = u_G(t) - p_1(0,t)/R_G, \tag{4-28}$$

que depois de fazermos as respectivas substituições a partir das equações (4-17), obtem-se

$$u_1^+(t) - u_1^-(t) = u_G(t) - \frac{\rho c}{R_G A_1}\left[u_1^+(t) - u_1^-(t)\right], \tag{4-29}$$

ou

$$u_1^+(t) = \frac{R_G}{R_G + \rho c/A_1} u_G(t) + \frac{R_G - \rho c/A_1}{R_G + \rho c/A_1} u_1^-(t), \tag{4-30}$$

de onde se deduz que o coeficiente de reflexão glotal é dado por

$$r_G = -\frac{R_G - \rho c/A_1}{R_G + \rho c/A_1} \tag{4-31}$$

e assim

$$u_1^+(t) = \frac{1 - r_G}{2} u_G(t) - r_G u_1^-(t). \tag{4-32}$$

Ficamos portanto finalmente na presença das relações que nos permitem estabelecer o diagrama de fluxo do modelo completo do sistema vocal quando representado por *N* tubos acústicos sem perdas — Figura 4-8.

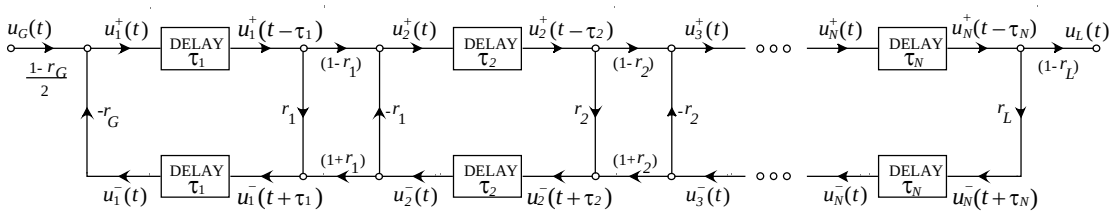


Figura 4-8: Diagrama de fluxo completo de um modelo de *N* tubos.

4.2.4 Conversão num Filtro Digital

De modo a podermos implementar o modelo acústico por intermédio de um filtro digital, necessitamos de converter o modelo analógico obtido num sistema discreto no tempo. Para isso é conveniente impormos que todos os tubos que compõem o modelo possuam o mesmo comprimento. Assim, se o modelo for formado por *N* tubos e possuir um tamanho total *l*, cada um dos tubos terá um comprimento l/N; isto é, $l_k = l/N$ para *k* = 1, 2, ..., *N*, representando l_k o comprimento do *k* ésimo tubo. E portanto o tempo gasto por uma onda progressiva a percorrer cada um dos tubos será o mesmo e igual a $\tau = l/(cN)$.

Se aplicarmos na entrada um impulso unitário, $u_G(t) = \delta(t)$, obtemos à saída a resposta impulsional do sistema. Como sabemos, sempre que um dado impulso atinge uma junção entre dois tubos, é decomposto em outros dois impulsos, um dos quais é transmitido para o tubo seguinte e o outro é reflectido, invertendo dessa forma a sua

progressão. Assim, ao longo das várias junções entre os tubos colocados em série, os impulsos vão-se duplicando sucessivamente. Só após $N\tau$ segundos é que uma pequena parcela do impulso inicialmente aplicado na entrada do sistema atinge a saída. Parcela resultante das sucessivas atenuações impostas pela propagação através de todas as junções uma só vez, que pode ser representada por $\alpha_0\delta(t - N\tau)$. Sempre que ocorra uma reflexão de um qualquer impulso que se dirija em direcção à saída do sistema, a parcela reflectida sofrerá um atraso de 2τ segundos gastos para se deslocar em ambas as direcções de um único tubo, ou então atrasos múltiplos de 2τ segundos no caso de recuar mais do que um tubo. Assim, após o impulso $\alpha_0\delta(t - N\tau)$, atingirão a saída outros impulsos desfasados deste primeiro, mas sempre com atrasos múltiplos de 2τ segundos, e portanto a resposta impulsional do sistema terá obrigatoriamente a seguinte forma

$$\begin{aligned} v_a(t) &= \alpha_0\delta(t - N\tau) + \alpha_1\delta(t - N\tau - 2\tau) + \alpha_2\delta(t - N\tau - 4\tau) + \ldots \\ &= \sum_{k=0}^{+\infty} \alpha_k \delta(t - N\tau - 2k\tau) \end{aligned} \tag{4-33}$$

Uma vez que a resposta do sistema a um impulso unitário é constituída por uma série de impulsos uniformemente espaçados de 2τ segundos, se aplicarmos na entrada um sinal amostrado com um período de amostragem 2τ — obviamente este sinal não amostrado terá de ser limitado na frequência a valores inferiores a $\pi/(2\tau)$ —, o modelo comportar-se-á como um sistema discreto, podendo por isso ser implementado por um filtro digital cuja resposta impulsional, no caso de o modelo ser constituido por um número par de tubos, será dado por

$$v(n) = \begin{cases} 0 & n < N/2 \\ \alpha_{n-N/2} & n \geq N/2 \end{cases} \tag{4-34}$$

ou seja, $v(n) = v_a(nT)$, com $T = 2\tau$ representando o período de amostragem.

Note-se que o N não tem que ser necessariamente par; no entanto a sua consideração como tal é admitida uma vez que se o não fosse, teria que se utilizar interpolação para se obter as amostras de saída. Assim o diagrama de fluxo do sistema discreto no tempo correspondente é o que se representa na Figura 4-9.

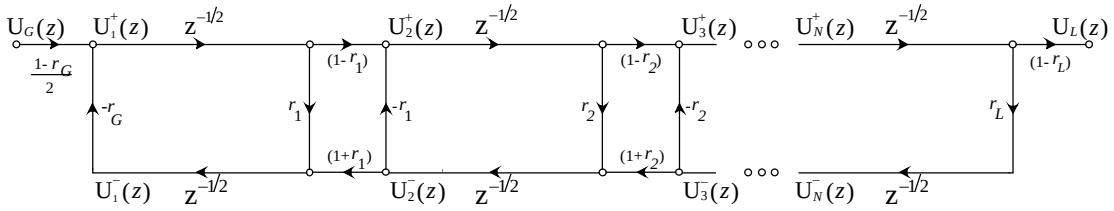


Figura 4-9: Diagrama de fluxo do sistema discreto do modelo de *N* tubos.

Na conversão do diagrama de fluxo do sistema analógico no diagrama de fluxo para o sistema discreto, cada atraso de propagação τ, correspondendo a meio período de amostragem, passa a ser representado pela transmitância $z^{-1/2}$.

Estamos finalmente interessados em derivar uma expressão geral para a função de transferência do filtro em termos dos coeficientes de reflexão $\{r_k\}$. Pretende-se portanto encontrar a seguinte função de transferência

$$V(z) = U_L(z)/U_G(z). \tag{4-35}$$

Para isso, vamos expressar $U_G(z)$ em função de $U_L(z)$, e não ao contrário, uma vez que não há possibilidade de expressarmos $U_1^+(z)$ e $U_1^-(z)$ em função da excitação $U_G(z)$.

Da Figura 4-9 podemos retirar as seguintes relações em transformadas de z referentes a cada junção entre dois tubos

$$U_{k+1}^+(z) = (1-r_k)z^{-1/2}U_k^+(z) - r_k U_{k+1}^-(z), \tag{4-36a}$$

$$U_k^-(z) = r_k z^{-1}U_k^+(z) + (1+r_k)z^{-1/2}U_{k+1}^-(z). \tag{4-36b}$$

De seguida, resolvendo as equações para $U_k^+(z)$ e $U_k^-(z)$, obtem-se

$$U_k^+(z) = \frac{z^{1/2}}{1-r_k}U_{k+1}^+(z) + \frac{r_k z^{1/2}}{1-r_k}U_{k+1}^-(z), \tag{4-37a}$$

$$U_k^-(z) = \frac{r_k z^{-1/2}}{1-r_k}U_{k+1}^+(z) + \frac{z^{-1/2}}{1-r_k}U_{k+1}^-(z). \tag{4-37b}$$

Se considerarmos o vector $\mathbf{U}_k$ dado por

$$\mathbf{U}_k = \begin{bmatrix} U_k^+(z) \\ U_k^-(z) \end{bmatrix} \tag{4-38}$$

e a matriz $\mathbf{Q}_k$ dada por

$$\mathbf{Q}_k = \begin{bmatrix} \dfrac{z^{1/2}}{1-r_k} & \dfrac{r_k z^{1/2}}{1-r_k} \\ \dfrac{r_k z^{-1/2}}{1-r_k} & \dfrac{z^{-1/2}}{1-r_k} \end{bmatrix} \tag{4-39}$$

as equações (4-37) podem ser expressas de forma mais compacta através da seguinte relação matricial

$$\mathbf{U}_k = \mathbf{Q}_k \mathbf{U}_{k+1} \tag{4-40}$$

Deste modo, através de um produtório matricial, facilmente identificamos a relação existente entre o vector dos fluxos à entrada do primeiro tubo e o correspondente vector referente ao último tubo do modelo,

$$\mathbf{U}_1 = \mathbf{Q}_1 \mathbf{Q}_2 \ldots \mathbf{Q}_{N-1} \mathbf{U}_N \,. \tag{4-41}$$

Sendo possível expressar o valor da transformada do fluxo à entrada do sistema, $\mathrm{U}_G(z)$, em função do vector $\mathbf{U}_1$

$$\mathrm{U}_G(z) = \frac{2}{1-r_G}\mathrm{U}_1^+(z) + \frac{2r_G}{1-r_G}\mathrm{U}_1^-(z) = \begin{bmatrix} \dfrac{2}{1-r_G} & \dfrac{2r_G}{1-r_G} \end{bmatrix} \mathbf{U}_1 \,, \tag{4-42}$$

bem como o vector $\mathbf{U}_N$ em função da transformada do fluxo à saída do sistema, $\mathrm{U}_L(z)$,

$$\mathbf{U}_N = \begin{bmatrix} \dfrac{z^{1/2}}{1-r_L}\mathrm{U}_L(z) \\ \dfrac{r_L z^{-1/2}}{1-r_L}\mathrm{U}_L(z) \end{bmatrix} = \mathbf{Q}_N \begin{bmatrix} 1 \\ 0 \end{bmatrix} \mathrm{U}_L(z), \quad (\textit{admitindo } r_N \equiv r_L) \tag{4-43}$$

encontramos finalmente a relação pretendida,

$$\mathrm{U}_G(z) = \begin{bmatrix} \dfrac{2}{1-r_G} & \dfrac{2r_G}{1-r_G} \end{bmatrix} \mathbf{Q}_1 \mathbf{Q}_2 \ldots \mathbf{Q}_{N-1} \mathbf{Q}_N \begin{bmatrix} 1 \\ 0 \end{bmatrix} \mathrm{U}_L(z) \,. \tag{4-44}$$

Colocando a transmitância $z^{1/2}$ em evidência, a matriz $\mathbf{Q}_k$ dá origem a uma outra matriz, $\overline{\mathbf{Q}}_k$, constituída apenas por elementos constantes ou proporcionais a z^{-1}, isto é

$$\mathbf{Q}_k = z^{1/2} \begin{bmatrix} \dfrac{1}{1-r_k} & \dfrac{r_k}{1-r_k} \\ \dfrac{r_k z^{-1}}{1-r_k} & \dfrac{z^{-1}}{1-r_k} \end{bmatrix} = z^{1/2} \overline{\mathbf{Q}}_k \,. \tag{4-45}$$

Podemos, portanto expressar o inverso da função de transferência do filtro da seguinte forma

$$\frac{1}{\mathrm{V}(z)} = z^{N/2}\begin{bmatrix} \dfrac{2}{1-r_G} & \dfrac{2r_G}{1-r_G} \end{bmatrix}\overline{\mathbf{Q}}_1\overline{\mathbf{Q}}_2\ldots\overline{\mathbf{Q}}_{N-1}\overline{\mathbf{Q}}_N\begin{bmatrix}1\\0\end{bmatrix}. \tag{4-46}$$

Sendo cada matriz $\overline{\mathbf{Q}}_k$ formada apenas por termos constantes e termos proporcionais a z^{-1}, facilmente se conclui que o produto das várias matrizes na equação (4-46) resulta num polinómio de variável z^{-1} de grau *N*. O termo $z^{N/2}$ corresponde a um atraso de $N/2$ amostras. Portanto, a função de transferência de um modelo formado por *N* tubos acústicos sem perdas terá o seguinte aspecto

$$\mathrm{V}(z) = \frac{Kz^{-N/2}}{\mathrm{D}_N(z)}, \tag{4-47a}$$

onde a constante *K* é dada por

$$K = 0.5(1-r_G)(1-r_1)(1-r_2)\ldots(1-r_{N-1})(1-r_L), \tag{4-47b}$$

e o denominador por

$$\mathrm{D}_N(z) = \begin{bmatrix}1 & r_G\end{bmatrix}\begin{bmatrix}1 & r_1\\ r_1z^{-1} & z^{-1}\end{bmatrix}\ldots\begin{bmatrix}1 & r_{N-1}\\ r_{N-1}z^{-1} & z^{-1}\end{bmatrix}\begin{bmatrix}1 & r_L\\ r_Lz^{-1} & z^{-1}\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} \tag{4-47c}$$

isto é, tem a forma de um polinómio em z^{-1} de grau *N*,

$$\mathrm{D}_N(z) = 1-\alpha_1z^{-1}-\alpha_2z^{-2}\cdots-\alpha_Nz^{-N}, \tag{4-48}$$

o que está em concordância com as deduções feitas anteriormente quando se concluiu que o sistema possuía uma resposta impulsional infinita com um atraso $N\tau = N/2\cdot T$ segundos — equações (4-33) e (4-34). Adicionalmente ficamos também a saber através da equação (4-47a) que a função de transferência não possui zeros. É constituída apenas por pólos que, como sabemos, estão relacionados com as ressonâncias que caracterizam as formantes do tracto vocal.

No caso de considerarmos a impedância Z_G infinitamente grande ($r_G = -1$) o polinómio $\mathrm{D}_N(z)$ pode ser obtido recursivamente a partir da equação (4-47c) [Rabiner (79)]. Mesmo admitindo $r_G = -1$, devemos decidir ainda sobre o tipo de carga de radiação com que o modelo termina. Tal como se depreende da Figura 4-9, a carga de radiação pode ser interpretada como um tubo adicional de tamanho infinito à saída do modelo, uma vez que não possui qualquer onda reflectida. Esta é a única fonte de perdas se supusermos $r_G = -1$, e por isso a escolha da área deste tubo imaginário é determinante na largura de banda das ressonâncias de $\mathrm{V}(z)$. Assim,

normalmente o valor atribuído à sua área será o adequado a se obter o coeficiente de reflexão r_L que conduza a ressonâncias com larguras de banda apropriadas.

Como exemplo, admitindo o caso particular de o tubo ser de área infinita obtem-se $r_L = 1$, resultando assim um modelo sem perdas.

Finalmente, uma última questão que se levanta, é sobre o número de tubos que devem constituir o modelo. Como já referido, o presente modelo pode aproximar o comportamento do tracto vocal apenas na banda de frequências $-1/(2T) < f < 1/(2T)$, e o período de amostragem T encontra-se relacionado com o número de tubos N da seguinte forma: $T = 2\tau = 2l/(cN)$. Verifica-se, portanto, que a escolha do número de tubos, que determina a ordem do filtro digital correspondente, depende da taxa de amostragem utilizada na representação do sinal de voz. Se considerarmos um comprimento para o tracto vocal de l=17.5 cm e a velocidade do som dada por c=35000 cm/s, o número de tubos será então dado por $N = 1/(1000T)$. Assim, admitindo por exemplo uma frequência de amostragem de 8000 Hz — uma vez que o espectro de um sinal de voz possui a sua potência concentrada essencialmente em frequências inferiores a 4000 Hz —, obtêm-se $N = 8$ tubos. Este valor encontra-se em concordância com a constatação prática de que as ressonâncias do tracto vocal ocorrem com uma densidade de cerca de uma formante por cada 1000 Hz [Rabiner (79)]. Note-se que, sendo $N = 8$ o grau polinomial do denominador da função de transferência, existirão no máximo $N/2 = 4$ pólos conjugados para fornecerem as ressonâncias na banda de frequências $-4000\,\text{Hz} < f < 4000\,\text{Hz}$, isto é, obtem-se a densidade de uma formante por cada 1000 Hz.

A Figura 4-10 ilustra um exemplo da resposta na frequência dum modelo de 10 tubos ($N = 10$) para uma frequência de amostragem de 10 KHz.

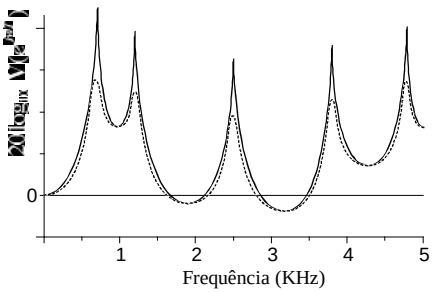


Figura 4-10: Resposta na frequência de um modelo formado por 10 tubos; curva a tracejado: modelo incluindo perdas; curva a cheio: modelo sem perdas.

Comparando a curva da resposta na frequência representativa do modelo sem perdas ($r_L = 1$) com a curva referente ao modelo com perdas ($0 < r_L < 1$), conclui-se que com uma apropriada carga de radiação labial obtem-se um modelo mais realista, (curva a tracejado) uma vez que a largura de banda das várias formantes deixa de ser nula.

4.2.5 Efeito das Perdas do Tracto Vocal

O modelo encontrado baseou-se no pressuposto de que não existia qualquer perdas ao longo do tubo acústico. Portanto, não se entrou em linha de conta com a energia dissipada através da fricção de viscosidade do ar nas paredes do tubo, nem tão pouco se consideraram as perdas relacionadas com a vibração e condutividade térmica das paredes. Uma vez que todas estas perdas são demasiado dependentes da frequência, a sua inclusão no modelo seria algo extremamente complexo. De qualquer forma, não deixa de ser instrutivo apontarmos que tipo de influências teriam todas essas perdas no comportamento do modelo resultante.

Devido à sua natureza elástica, as paredes do tracto vocal estão sujeitas a pequenas perturbações provocadas por variações de pressão do ar no seu interior, resultando por isso um tubo de secções transversais de área variáveis no tempo. Como consequência, as ressonâncias deixam de estar exactamente em cima do eixo *jw* no plano–s [Rabiner (79)], e por isso, as formantes passam a ter amplitudes finitas e larguras de banda não nulas. Para além disso, as formantes sofrem ainda um deslocamento para frequências ligeiramente superiores. Porém, todos estes efeitos fazem-se sentir de forma mais acentuada para baixas frequências, o que é compreensível uma vez que não seria de esperar grandes movimentos das paredes em resposta a perturbações de alta frequência.

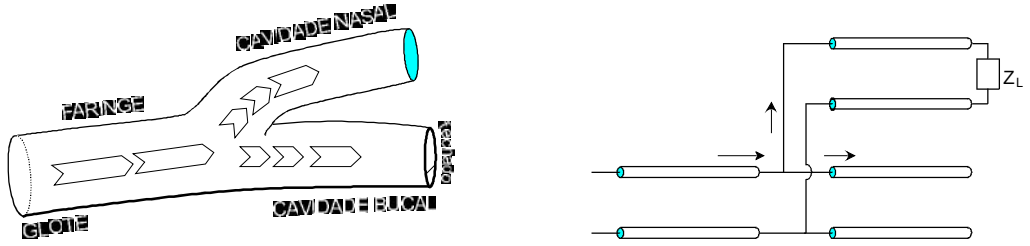
Embora as perdas de condutividade térmica e de fricção de viscosidade contribuam para aumentar ainda mais a largura de banda das formantes, contrariamente ao efeito anterior, incluindo este tipo de perdas as formantes deslocam-se para frequências ligeiramente inferiores, e são as de frequências mais elevadas que ficam sujeitas a maior atenuação. Globalmente, estas perdas tem um efeito menos acentuado do que o referente às perdas por vibração.

Uma vez que para frequências inferiores a 3-4 KHz o efeito das perdas térmicas e de fricção é pequeno quando comparado com o efeito das perdas de vibração, as formantes no modelo final encontram-se deslocadas para frequências ligeiramente superiores em relação ao modelo sem perdas, e são as formantes inferiores que são sujeitas a maior atenuação.

De uma forma resumida, podemos dizer que a largura de banda da primeira formante é fundamentalmente determinada pelas perdas relacionadas com a natureza elástica das paredes do tracto vocal, enquanto que a largura de banda das formantes mais elevadas são determinadas essencialmente pelas perdas de radiação labial. A largura de banda da segunda e terceira formantes são determinadas pela combinação destes dois tipos de perdas. Podemos ainda considerar alguma influência das perdas glotais na largura de banda das formantes mais baixas, bem como a influência das perdas térmicas e de fricção na largura de banda das formantes mais elevadas.

4.3 Efeito do Acoplamento Nasal

O tracto vocal pode ser caracterizado, como já vimos, por um conjunto de tubos acústicos de diferentes secções dispostos em série. A estes ainda se poderá associar um outro em paralelo, como ilustrado na Figura 4-11a, representando o acoplamento do tracto nasal.



a) b)

Figura 4-11: Modelos para a produção de sons nasais: a) Tubo acústico; b) Linha de transmissão.

Na produção de sons nasais, a cavidade nasal tem um papel preponderante no tipo de sons obtidos. Porém, há que distinguir dois processos de produção distintos, responsáveis pela obtenção dos sons nasais existentes: consoantes nasais e vogais nasais.

Na produção das consoantes nasais — /m/, /n/ e /nh/ — o véu palatino encontra-se abaixado de modo a acoplar a cavidade nasal com o tracto vocal; e em simultâneo, existe uma obstrução completa num ponto da cavidade bucal. Tal como se ilustra na Figura 4-11a, esta configuração possui dois ramais, encontrando-se um deles fechado. Refira-se ainda que no ponto de bifurcação a pressão é a mesma à entrada de cada ramal, e o fluxo nesse ponto é igual à soma das fluxos à entrada das cavidades bucal e nasal. Assumindo que a pressão e o fluxo têm como análogos no modelo eléctrico, respectivamente o potencial e a corrente eléctrica, a linha de transmissão da Figura 4-11b representa o equivalente eléctrico do modelo acústico. Repare-se que nos sons consonânticos nasais, a radiação do som ocorre primeiramente através saída nasal (cavidade bucal fechada). Por essa razão, no equivalente eléctrico, a linha de transmissão correspondente ao tracto nasal termina numa impedância de radiação apropriada ao tamanho da abertura das narinas, enquanto que a linha referente ao tracto oral termina em circuito aberto, uma vez que não existe qualquer fluxo nessa extremidade.

Na produção das vogais nasais a configuração do tracto vocal é semelhante ao utilizado na produção das vogais orais. A diferença consiste basicamente de que nas primeiras o som resultante deriva da sobreposição das saídas orais e nasais (modelos da Figura 4-11 com a cavidade bucal aberta).

Em semelhança com os sons não nasais, também se poderia derivar um modelo matemático para este tipo de configuração. Contudo, uma vez que a função de transferência obtida teria muitas características comuns com o modelo matemático anterior, é possível prever de algum modo a função de transferência da nova configuração, sem necessidade de recorrermos a complicadas equações. Assim, o sistema será caracterizado por uma série de ressonâncias ou formantes que serão dependentes, quer da forma, quer do comprimento de três tubos: faríngeo, bucal e nasal. A diferença mais significativa em relação ao modelo de sons não vozeados reside no facto de que a cavidade bucal fechada pode capturar a energia de certas frequências, evitando que atinjam a saída via cavidade nasal. É um efeito semelhante ao que se obteria no caso de, para algumas frequências, a junção — ou ponto de bifurcação — ser curto-circuitada pela linha de transmissão que representa a cavidade bucal. Por isso, para sons nasais, a função de transferência do sistema será

caracterizada, não apenas por ressonâncias, mas também por zeros ou anti-ressonâncias. Verifica-se ainda que as formantes nasais possuem maior "largura de banda" do que as não nasais. Isto é atribuído à maior fricção de viscosidade e perdas de condutividade térmica devido à maior área das paredes da cavidade nasal [Rabiner (79)].

4.4 Modelo Final

Nas secções anteriores abordamos algumas das características básicas de um sinal de voz e mostramos de que modo podem ser relacionadas com o processo de produção acústica. Basicamente vimos que a voz é gerada essencialmente a partir de dois tipos de excitação, sendo cada um deles responsável por um som distinto à saída do sistema acústico. O som resultante depende igualmente do tipo de ressonâncias a que a excitação é sujeita ao passar pelo tracto vocal — Figura 4-12. Assim, um modelo para representar todo o processo de produção de voz tem que entrar em linha de conta com estes aspectos.

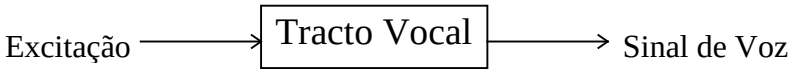


Figura 4-12: Diagrama de blocos do modelo simplificado de produção de voz.

Estamos particularmente interessados que o modelo encontrado e o sistema físico de produção tenham saídas equivalentes. Assim, o modelo discreto baseado em tubos acústicos sem perdas, encontrado anteriormente, pode ser utilizado com esse propósito, pois as suas características são apropriadas para o sistema pretendido. Recorde-se que para a sua formulação baseamo-nos nas características físicas do tracto vocal, e sendo caracterizado por coeficientes de reflexão, ou equivalentemente por um conjunto de áreas, através da manipulação desses parâmetros o sistema pode ser controlado de modo a se obter um sinal de saída com propriedades equivalentes às de um sinal de voz. Neste caso, servindo-nos da função de transferência expressa nas equações (4-47), o modelo será constituido por um sistema linear tal como ilustrado na Figura 4-13.

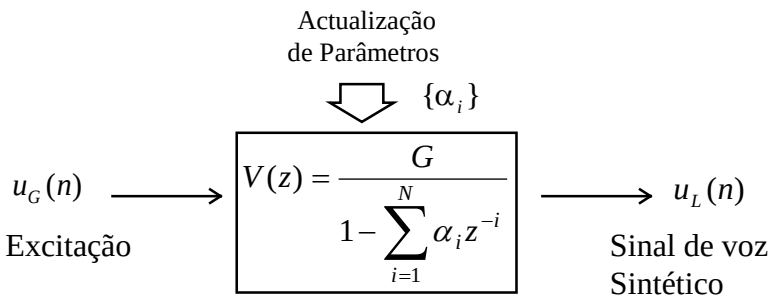


Figura 4-13: Representação do tracto vocal com base no modelo dos tubos sem perdas.

Uma vez que o atraso *N*/2 incluído no numerador da equação (4-47a) não tem qualquer efeito nas características ressonantes do sistema, optou-se por o suprimir na função de transferência do modelo. Recorde-se que tanto o ganho *G* como os coeficientes $\{\alpha_i\}$ da função de transferência *V*(*z*) se encontram intimamente relacionados com a área da secção transversal dos vários tubos usados no modelo.

De modo a haver uma permanente similaridade entre o sinal sintetizado e o sinal de voz, o tipo de excitação e as características ressonantes do sistema devem ser actualizadas ao longo do tempo. Felizmente constata-se que as propriedades de um sinal de voz alteram-se lentamente. Por isso, para uma grande parte dos sons vocais é possível assumir que as principais propriedades de excitação e do tracto vocal permanecem fixas para intervalos de tempo na ordem dos 10 a 20 ms, isto é, considera-se que o sinal de voz é aproximadamente estacionário para curtos espaços de tempo. Dessa forma, adaptando os parâmetros do sistema digital periodicamente, a uma cadência correspondente ao intervalo de tempo de estacionaridade, consegue-se uma boa aproximação com o sistema acústico humano. Do mesmo modo, o tipo de excitação será controlado periodicamente, comutando entre um sinal formado por impulsos quase periódicos, para sons vozeados; e um constituido por ruído aleatório, para sons não vozeados. Isto porque a maioria dos sons fonéticos podem ser classificados em vozeados e não vozeados.

Este modelo formado apenas por pólos é suficiente para representar o efeito do tracto vocal para a maioria dos sons vocais; no entanto os sons nasais requerem, para além de ressonâncias, também a existência de zeros ou anti-ressonâncias. Por isso, de modo a obtermos um modelo mais completo devemos incluir zeros na função de transferência ou então, alternativamente, o mesmo efeito poderá ser conseguido através do acréscimo do número de pólos, uma vez que, como é do conhecimento

geral, os zeros espectrais podem sempre ser representados por um número suficientemente grande de pólos. Esta segunda opção é preferível. Pois, se o tracto vocal for modelado por uma função de transferência contendo apenas pólos, dispomos de vários algoritmos de predição linear, de reconhecida eficácia na estimação dos parâmetros do modelo.

Uma vez que as áreas dos tubos são todas positivas, é fácil mostrar que os respectivos coeficientes de reflexão estão limitados por

$$-1 \le r_k \le 1 \tag{4-49}$$

e nesta situação é também possível demonstrar que todos os pólos de $V(z)$ se encontram dentro do circulo unitário, ficando assim garantida a estabilidade do sistema.

Repare-se contudo que a função $V(z)$ relaciona apenas o fluxo de excitação com o fluxo à saída do sistema (lábios). Para se obter a pressão acústica à saída, como é normalmente o pretendido, deve ser incluído o efeito da carga de radiação. Como referido na secção 4.2.2 a pressão acústica e o fluxo nos lábios encontram-se relacionadas através de uma impedância de radiação, $\mathrm{Z}_L(z)$. Dado que

$$\mathrm{U}_L(z) = V(z)\mathrm{U}_G(z), \tag{4-50}$$

após convertermos a equação (4-12) na sua equivalente para o modelo discreto, é possível obter a partir destas duas relações a seguinte pressão acústica à saída do sistema,

$$\mathrm{P}_L(z) = \mathrm{Z}_L(z)V(z)\mathrm{U}_G(z). \tag{4-51}$$

A partir do que foi referido na secção 4.2.2 deduz-se que ao nível dos lábios a pressão se encontra relacionada com o fluxo por uma operação do tipo filtragem passa-baixo, podendo mesmo considerar-se que para baixas frequências a pressão pode ser aproximada pela derivada do respectivo fluxo. Por isso o efeito da radiação pode ser razoavelmente aproximado por

$$\mathrm{Z}_L(z) = R_0(1 - z^{-1}). \tag{4-52}$$

Para o modelo ficar completo falta ainda referirmos o modo como a excitação pode ser gerada. Para os sons não vozeados resume-se a uma fonte de ruído aleatório e a um ganho de controle de intensidade. Para os sons vozeados a forma de onda do sinal de excitação deve ter alguma semelhança com os impulsos glotais da Figura 4-2. Por isso, um possível modelo para este tipo de excitação inclui uma fonte de impulsos

espaçados pelo período fundamental (*pitch*) que alimenta um sistema linear de resposta impulsional $g(n)$ com a forma semelhante à do impulso glotal. Uma forma possível de representar este impulso é através da função ilustrada na Figura 4-14, dada por

$$g(n) = \begin{cases} 0.5\left[1 - \cos(\pi n / N_1)\right] & 0 \le n \le N_1 \\ \cos(\pi(n - N_1)/2N_2) & N_1 \le n \le N_1 + N_2 \\ 0 & \text{restantes} \end{cases} \tag{4-53}$$

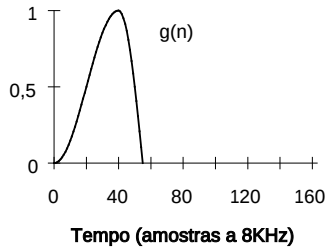


Figura 4-14: Forma aproximada do impulso glotal.

O efeito desta função é introduzir uma filtragem do tipo passa-baixo. Uma vez que a função em causa tem suporte temporal finito, a sua transformada de *z*, $G(z)$, possui apenas zeros. No entanto, modelos do tipo AR podem igualmente ser usados com sucesso. Também neste tipo de excitação deverá existir um ganho que controle a intensidade.

Concluindo, o modelo global deveria incluir todos os efeitos tal como mencionados; no entanto, como isso implicaria um sistema constituido por zeros, é conveniente combinar os modelos de excitação, do tracto vocal e de radiação num único sistema representado por uma função de transferência contendo apenas pólos, isto é

$$H(z) = G(z)V(z)\mathrm{Z}_L(z)\,. \tag{4-54}$$

A forma geral deste modelo digital de produção de voz encontra-se representada na Figura 4-15.

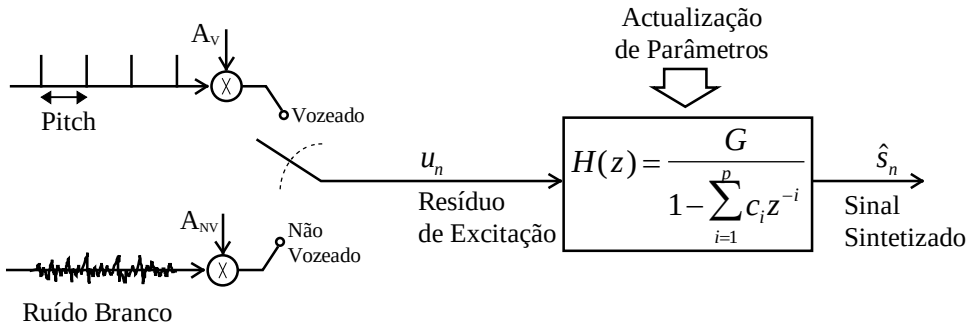


Figura 4-15: Modelo digital de produção de voz.

Note-se que, como $H(z)$ consiste numa função do tipo AR, podemo-nos servir de técnicas de Predição Linear para estimação dos seus parâmetros. O que diferencia esta função da função $V(z)$ — função que representa unicamente o tracto vocal — é a sua ordem $(p+1)$ que será necessariamente superior.

Embora este modelo se comporte bem para sons contínuos, como é o caso das vogais onde os parâmetros variam muito lentamente, terá necessariamente algumas limitações derivadas das várias aproximações e simplificações a que se recorreu para a sua formulação. De seguida apontamos algumas das mais importantes dessas limitações:

♦ Para sons transientes, como é o caso dos sons plosivos, o modelo não é adequado devido às transições abruptas das características espectrais desse tipo de sons;

♦ A não inclusão de zeros na função de transferência é consideravelmente limitativo para sons nasais, uma vez que a ordem do modelo AR não é normalmente suficiente — devido a necessidades de compactação e ao elevado peso computacional que implicaria uma ordem demasiado elevada — para compensar totalmente a ausência de zeros;

♦ A simples diferenciação da excitação em vozeada e não vozeada é inadequado para a produção de sons fricativos vozeados uma vez que nestes a fricção se encontra correlacionada com impulsos glotais;

♦ O modelo digital requer que os impulsos glotais sejam separados por inteiros, múltiplos do período de amostragem, não representando por isso a distância exacta do *pitch*. É possível no entanto contornar esta limitação usando técnicas de interpolação.

Apesar de todas estas imperfeições, felizmente nenhuma delas limita severamente a aplicabilidade do modelo. Mais relevante ainda, é a constatação de que o referido modelo representa o método mais utilizado na codificação de voz.

Capítulo 5

Algoritmos de Codificação

Neste capítulo apresenta-se uma abordagem geral dos vários algoritmos utilizados na codificação de voz. No entanto, uma descrição mais completa de todos os tipos de codificadores pode ser encontrada em [Spanias (94)].

5.1 Introdução

5.1.1 A Codificação

Tem-se assistido, sobretudo nesta última década, a um considerável progresso nas aplicações dos codificadores de voz de baixos ritmos de transmissão, tanto no que se refere a comunicações como em relação a aplicações de voz em computadores.

Embora a fibra óptica esteja a contribuir para que a largura de banda nas comunicações por fio seja já pouco dispendiosa, existe uma necessidade crescente para a conservação da largura de banda e para o aumento de privacidade nas comunicações celulares e por satélite. Existe também a necessidade da utilização racional de memória uma vez que, muitas das aplicações requerem armazenamento do sinal de voz no seu formato digital. Como um sinal de voz digital está associado a uma elevada quantidade de informação, e dessa forma, a exigentes larguras de banda de transmissão e requisitos de memória, a codificação de voz — responsável pela obtenção de representações compactas do sinal de voz — apresenta-se como o processo privilegiado no sentido de permitir uma eficiente utilização dos recursos, quer de transmissão quer de armazenamento. O objectivo da codificação de voz é representar a fala com um número mínimo de bits, mantendo contudo uma qualidade perceptual razoável no sinal reconstruído. Dependendo do modo de representação do sinal, existem essencialmente duas formas de abordar o problema da codificação.

A forma mais simples passa pela representação do sinal de voz por uma informação binária que permita a reconstrução directa da forma de onda no tempo. Os codificadores que se baseiam neste tipo de técnica, conhecidos por codificadores de forma de onda (*waveform coders*), permitem sinais reconstruídos de excelente qualidade. Limitam-se, no entanto, a taxas de compressão pouco ambiciosas.

A outra forma de abordar o problema da codificação, consiste em segmentar o sinal de voz em blocos (ou *frames*) relativamente pequenos (segmentos de 25 ms, por exemplo) e representar cada um deles por parâmetros que caracterizem um determinado modelo de produção de voz (codificadores de fonte ou *vocoders*), ou então por um conjunto de parâmetros espectrais (codificadores sinusoidais). Durante a análise (processo de codificação) o sinal é representado por uma série compacta de parâmetros que são quantificados eficientemente. No processo inverso (síntese) os parâmetros são descodificados e usados, em conjunto com o mecanismo de reconstrução, para formar a fala. Estes codificadores, designados por paramétricos, conduzem a sinais sintetizados de fraca qualidade — fala de qualidade sintética —, mas conseguem taxas de compressão extremamente elevadas.

Entre os dois tipos de codificadores mencionados — paramétricos e não paramétricos — existe uma série de codificadores a funcionarem a ritmos intermédios, que por combinarem características de ambas as técnicas de codificação, conseguem também conciliar as virtudes de ambas. Designadamente, aliam à eficiência de codificação reconhecida nos *vocoders* a excelente qualidade que caracteriza os codificadores de forma de onda. É por isso que estes codificadores "híbridos", também denominados por codificadores da terceira geração, constituem a metodologia de codificação mais adoptada actualmente para a compressão de voz, sempre que se pretende fala de boa qualidade a baixos ritmos de transmissão. Este bom desempenho é conseguido fundamentalmente à custa do aumento da complexidade do algoritmo de codificação. A maior parte destes codificadores, para além de incluírem mecanismos para representação das propriedades espectrais da fala bem como meios que garantam uma aproximação com a forma de onda original, incluem também processos de optimização de desempenho dependentes das características perceptuais do ouvido humano.

As várias metodologias de codificação permitem ritmos de transmissão desde os 64 Kbps — codificadores de forma de onda — até algumas centenas de bits por segundo — codificadores de fonte. Os 64 Kbps é considerado o ritmo de transmissão não comprimido e deriva da técnica de codificação mais simples, o codificador standard PCM logarítmico, que consiste unicamente na quantificação directa da amplitude das amostras. Uma vez que se assume não haver compressão, este codificador é normalmente tido como referência para comparações de desempenho com os restantes codificadores.

5.1.2 Medidas de Avaliação de Desempenho

De um modo geral o desempenho de um dado algoritmo de codificação é avaliado com base em cinco parâmetros: qualidade do sinal reconstruído, factor de compressão, complexidade do algoritmo, atraso introduzido e robustez do algoritmo.

A robustez do algoritmo subentende protecção contra erros de canal e interferências acústicas ou, pelo menos, pouca sensibilidade a esses efeitos. Codificadores robustos, normalmente incluem algoritmos de correcção de erros para proteger toda a informação perceptualmente importante contra erros de canal. Além disso, em algumas aplicações, os codificadores terão de conseguir um razoável desempenho sobre uma diversidade de linguagens, ou quando a fala se encontra corrompida com ruído branco.

O atraso pode-se tornar num aspecto importante, fundamentalmente quando se pretende estabelecer diálogo. O atraso resultante do processo de codificação mais o de descodificação, introduzido pelos algoritmos híbridos para baixos ritmos de transmissão, situa-se normalmente entre os 50 e os 60 ms.

A complexidade do algoritmo de codificação é uma característica também a ter em conta, principalmente quando a complexidade se traduz num elevado peso computacional. Por exemplo, implementações em tempo-real de algoritmos híbridos para baixos ritmos de transmissão requerem tipicamente uma DSP capaz de executar acima de 12 MIPS (milhões de instruções por segundo).

O factor de compressão mede precisamente o grau de eficácia do codificador na realização daquela que é a sua principal missão — a de comprimir. Actualmente, praticamente todas as pesquisas de codificação subentendem taxas de compressão que

conduzam a ritmos de transmissão inferiores a 16 Kbps, pois mesmo a estes ritmos, os actuais processos de codificação conseguem manter o sinal codificado com boa qualidade perceptual. Os ritmos de transmissão inferiores a 16 Kbps podem ser classificamos em: *médios ritmos*, quando situados entre os 8 e os 16 Kbps; *baixos ritmos*, quando situados entre 2.4 e 8 Kbps; e *ritmos extremamente baixos*, quando inferiores a 2.4 Kbps.

Nas comunicações digitais a qualidade da fala é normalmente classificada em quatro categorias: qualidade de comentário, de rede, de comunicação e sintética. A fala com qualidade de "comentário" refere-se a fala de grande qualidade, conseguida quase sempre apenas a ritmos superiores a 64 Kbps. A qualidade de rede refere-se a uma qualidade comparável à de um sinal de voz analógico convencional, limitado na frequência entre 200 e 3200 Hz — de um modo geral as comunicações digitais de voz pressupõem que o sinal de voz seja limitado na frequência a 3.2 ou a 4 KHz e amostrado a 8 KHz; por isso, a não ser que algo seja dito em contrário, assumiremos sempre ser esta a frequência de amostragem de um sinal de voz —, e pode ser obtida com ritmos a partir de 16 Kbps. A qualidade de comunicação implica por vezes uma qualidade de fala algo degradada, que soa contudo natural e é altamente inteligível, sendo por isso adequada para telecomunicações, podendo ser obtida com ritmos a partir dos 4.8 Kbps. Por fim, a qualidade sintética refere-se a fala geralmente inteligível mas que soa pouco natural e à qual está associada a perda do reconhecimento do falante. Um dos objectivos da codificação actual, é conseguir qualidade de comunicação apenas a 4 Kbps uma vez que, por enquanto, os codificadores que operam até aos 4 Kbps limitam-se a produzir fala de qualidade sintética.

A medição da qualidade da fala é uma importante mas também extremamente difícil tarefa. A relação-sinal-ruído (SNR) é a medida objectiva mais utilizada na avaliação do desempenho dum algoritmo de compressão. Esta é dada por

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1}\left(s(n)-\hat{s}(n)\right)^2}\right), \qquad (5\text{-}1)$$

sendo $s(n)$ o sinal de voz original e $\hat{s}(n)$ o sinal de voz reconstruído. Contudo, esta medida é mais sensível ao ruído de reconstrução nas zonas de maior potência de sinal, dando "pouca importância" às zonas de baixa potência. Por isso, as variações de desempenho podem ser melhor quantificadas através duma média de relações sinal-ruído localizadas. Isto é, calculando a SNR para cada segmento de N amostras (SNR localizada), a SNR segmental dada por

$$\mathrm{SNR}_{\mathrm{SEG}} = \frac{10}{L}\sum_{i=0}^{L-1}\log_{10}\left(\frac{\sum_{n=iN}^{(i+1)N-1} s^2(n)}{\sum_{n=iN}^{(i+1)N-1}\left(s(n)-\hat{s}(n)\right)^2}\right) \tag{5-2}$$

será uma medida de desempenho mais representativa, até porque, sendo feita a média após o logaritmo, a SNR segmental penaliza sobretudo os codificadores de desempenho variável. Existem ainda outras possíveis medidas objectivas de desempenho, mas todas elas — incluindo as duas já referidas — além de serem quase sempre sensíveis a variações de ganho e a atrasos, tipicamente não entram em linha de conta com as características perceptuais do ouvido humano. Assim, tornam-se necessárias avaliações de caracter subjectivo, até porque muitos dos codificadores para baixos ritmos baseiam-se em critérios perceptuais.

O MOS (*Mean Opinion Score*) é um dos mais populares testes subjectivos. Usualmente envolve 12 a 24 ouvintes que são instruídos a classificar sinais de teste de acordo com uma escala qualitativa de cinco níveis: má, pobre, sofrível, boa ou excelente. Para que a fala possa ser classificada de qualidade "excelente" é necessário que, para além da ausência de qualquer ruído perceptível, seja indistinguível da versão não codificada. Por outro lado, voz de "má" qualidade subentende a presença de ruído e artefactos extremamente incómodos. Os ouvintes são "calibrados", isto é, são familiarizados com as condições de audição e com a gama de qualidades de voz que vão encontrar. Com este método, as classificações são obtidas a partir da média dos valores numéricos atribuídos aos vários níveis qualitativos (má qualidade – 1; pobre – 2; sofrível – 3; boa – 4; excelente – 5). A classificação MOS encontra-se relacionada com a qualidade da fala do seguinte modo: uma classificação entre 4 a 4.5 MOS implica qualidade de rede, entre 3.5 e 4 implica qualidade de comunicação, e uma classificação entre 2.5 e 3.5 implica qualidade sintética. Importa ainda referir que a

classificação MOS — por ser subjectiva — pode diferir significativamente de uns testes para outros, não podendo por isso ser encarada como uma medida absoluta para comparação de codificadores. Além deste, podem-se ainda referir o DRT e o DAM [Tremain (93)] como outros dois exemplos de testes de qualidade subjectivos que se baseiam na opinião de vários ouvintes.

O método DRT (*Diagnostic Rhyme Test*) é uma medida de inteligibilidade onde a tarefa do ouvinte é reconhecer uma de entre duas palavras possíveis ao longo de uma série de pares rimantes (exs. *meat/heat* e *vault/fault*). Cada par rimante é escolhido de modo a que as consoantes iniciais difiram por um simples atributo fonético. A classificação final é dada pelo número de palavras correctamente detectadas menos o número de palavras incorrectas, a dividir pelo número total de palavras.

O método DAM (*Diagnostic Acceptability Measure*) é baseado nos resultados dos métodos de teste que classificam a qualidade de um sistema de comunicação em função da aceitabilidade da fala por um ouvinte treinado e normalizado. A base de dados para este método consiste em 12 sentenças por falante (três vozes masculinas e três femininas) pronunciadas a uma taxa de uma sentença por cada 4 s. Utiliza-se um PC para se obter respostas de um ouvinte a 9 distorções de sinal, 8 distorções de ruído de fundo e 3 efeitos globais, inteligibilidade, agradabilidade e aceitabilidade. Cada distorção é classificada numa escala de 0 (zero para distorção não detectada) a 9 (nove para distorção extremamente incómoda). A Tabela 5-1 mostra a relação existente entre as classificações segundo os vários critérios subjectivos de avaliação de desempenho.

Qualidade	MOS	DRT	DAM
Excelente	5	>96	>75
Boa	4	87-96	60-75
Sofrível	3	79-87	45-60
Pobre	2	70-79	30-45
Má	1	<70	<30

Tabela 5-1: Comparação de critérios de desempenho

Embora a qualidade do sinal reconstruído e o factor de compressão sejam de fundamental importância, o peso de cada um dos factores de desempenho depende essencialmente do tipo de aplicação.

5.2 Codificadores de Forma de Onda (*Waveform Coders*)

Os codificadores de forma de onda procuram uma representação da onda temporal, sem explorarem necessariamente o modelo de produção de voz subjacente. Estes codificadores normalmente são mais robustos do que os *vocoders*, na medida em que se comportam bem com uma grande variedade de sinais. A quantificação está sempre presente em qualquer codificador, e muitos dos codificadores de forma de onda utilizam a quantificação escalar e/ou vectorial como únicas técnicas de codificação. Existem ainda outros codificadores de forma de onda que exploram as redundâncias do sinal no domínio das transformadas.

5.2.1 Quantificação Escalar e Vectorial

A quantificação é utilizada, quer nos codificadores de forma de onda, quer nos codificadores paramétricos. Enquanto que nos primeiros são as amostras do sinal a serem quantificadas, nos outros é a representação paramétrica do sinal a quantificar. Os métodos de quantificação classificam-se em duas classes, designadas por quantificação escalar e quantificação vectorial.

❍ **Métodos de Quantificação Escalar**

Os métodos de quantificação escalar englobam os codificadores PCM (*Pulse Code Modulation*), DPCM (PCM Diferencial) e DM (Modulação Delta). O método PCM uniforme é um processo "sem memória" que quantifica as amplitudes por arredondamento de cada uma das amostras a um de uma série de valores discretos. Num PCM uniforme não adaptativo o tamanho do degrau — diferença entre níveis de quantificação adjacentes — é constante. Uma vez que este codificador não tem qualquer mecanismo de extracção de redundâncias do sinal, resulta no método de codificação mais simples, mas é também o que conduz a maiores ritmos de transmissão. O seu desempenho em termos de SNR pode ser estimado por [Jayant (84)]

$$\mathrm{SNR}_1 = 6B + K_1 \quad \text{(dB)}, \tag{5-3}$$

sendo B o número de bits por amostra, e K_1 um parâmetro dependente do tamanho do degrau de quantificação. O PCM não uniforme, como o próprio nome indica, utiliza

um degrau de tamanho variável. Normalmente utiliza um pequeno degrau para amplitudes que ocorram frequentemente, e um degrau maior nas amplitudes menos frequentes. O tamanho do degrau pode igualmente ser definido em função da forma da função densidade de probabilidade (FDP). Outra classe de codificadores PCM não uniformes são os PCM logarítmicos *μ-law* e *A-law*. Uma quantificação logarítmica de 7 bits consegue para um sinal de voz um desempenho idêntico a um quantificador uniforme de 12 bits. As variações da gama dinâmica podem ser exploradas através da utilização de um degrau adaptativo. É o caso do codificador APCM (PCM adaptativo).

O codificador DPCM (PCM diferencial) é um quantificador escalar mais eficiente do que os já mencionados, uma vez que utiliza a redundância do sinal de voz ao explorar a correlação existente entre amostras adjacentes. Na sua forma mais simples, o DPCM codifica apenas a diferença entre as sucessivas amostras, e o descodificador reconstroi o sinal por integração. Quando as diferenças entre as amostras adjacentes é quantificada apenas com 1 bit obtem-se o codificador DM (Modelação Delta) que representa, assim, uma subclasse da codificação DPCM. O tamanho do degrau no codificador DM pode ainda adaptar-se ao longo do tempo com base nas propriedades estatísticas do sinal, dando origem ao codificador ADM (DM adaptativo). Os codificadores DM e DPCM são codificadores de complexidade média/baixa e comportam-se melhor do que o PCM para taxas inferiores a 32 Kbps. O codificador ADPCM tem-se mostrado versátil em aplicações de voz; foi inclusive adoptado pela CCITT para o standard G.721 a funcionar a 32 Kbps. O desempenho conseguido por este codificador, em termos de escala MOS, situa-se acima dos 4 valores. O algoritmo G.721 foi ainda modificado de modo a acomodar ritmos de 24 e 40 Kbps no standard G.727. Constata-se que o desempenho do ADPCM degrada-se rapidamente para taxas inferiores a 24 Kbps.

❍ **Métodos de Quantificação Vectorial**

A compressão de dados via quantificação vectorial (QV) é conseguida por codificação de uma série de amostras na forma de bloco ou vector. Embora desde cedo se tenha chegado à conclusão, através de estudos sobre distorção, ser possível atingir melhor desempenho com a QV do que o conseguido pela quantificação escalar; devido a toda a complexidade inerente àquela técnica, só recentemente se tem obtido

resultados práticos com a QV. Em particular, devido ao aparecimento de métodos eficientes de manipulação de blocos de dados de elevada dimensionalidade, a QV encontra-se actualmente associada à codificação de boa qualidade para baixos ritmos.

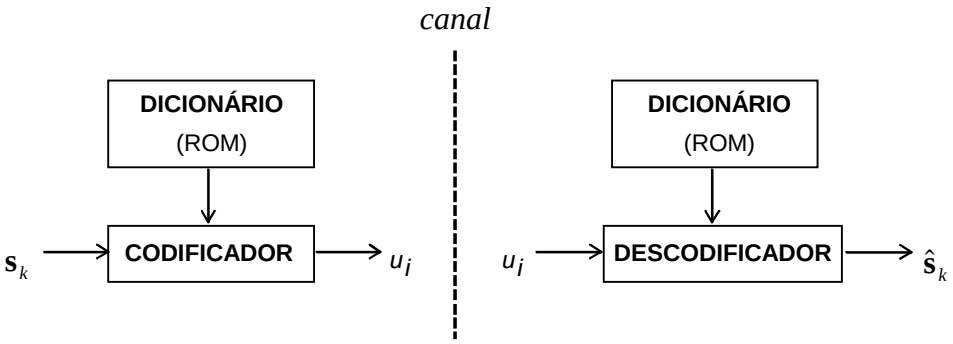


Figura 5-1: Diagrama de blocos do processo de quantificação vectorial.

A QV encontra-se genericamente representada pelo esquema da Figura 5-1. É constituída por um quantificador *N*-dimensional e um "dicionário" formado por um conjunto fixo de vectores, também *N*-dimensionais, que passaremos a designar por "palavras de código". Os vectores de entrada $\{\mathbf{s}_k\}$ são formados por amostras consecutivas do sinal que pretendemos quantificar, *s*(*n*). O quantificador faz corresponder ao *k* ésimo vector de entrada, $\mathbf{s}_k = \left[s_k(0) \quad s_k(1) \quad \cdots \quad s_k(N-1)\right]^T$, um índice, $\{u_i, i = 1,2,\cdots,L\}$, que identifique uma determinada palavra de código. O dicionário é constituido por *L* palavras de código $\{\hat{\mathbf{s}}_i = \left[\hat{s}_i(0) \quad \hat{s}_i(1) \quad \cdots \quad \hat{s}_i(N-1)\right]^T$ com $i = 1,2,\cdots,L\}$ que residem em memória tanto no codificador como no descodificador. Na procura da palavra de código, o codificador compara o vector de entrada, $\mathbf{s}_k$, com cada uma das palavras do dicionário, e o endereço da palavra de código que melhor se aproxime do vector de entrada, tendo em consideração uma medida de distorção ou um critério de verosimilhança $\varepsilon(\mathbf{s}_k, \hat{\mathbf{s}}_i)$, determina o índice da posição relativa no dicionário do vector seleccionado. A medida de distorção mais utilizada é a distância Euclidiana, dada pela soma dos erros quadráticos, ou seja

$$\varepsilon(\mathbf{s}_k, \hat{\mathbf{s}}_i) = \left\|\mathbf{s}_k - \hat{\mathbf{s}}_i\right\|^2 = \sum_{n=0}^{N-1}\left(s_k(n) - \hat{s}_i(n)\right)^2 . \tag{5-4}$$

Os *L* vectores (palavras de código) do dicionário são projectados dividindo o espaço vectorial em *L* células não sobrepostas, tal como ilustrado na Figura 5-2. Cada

célula C_i fica associada a um vector $\hat{\mathbf{s}}_i$ que funciona como centroide. Assim, o quantificador atribui o índice u_i ao vector $\mathbf{s}_k$ se este se encontrar no interior da célula C_i. Isto significa que se $\mathbf{s}_k$ se encontrar dentro da célula C_i, então será usado no descodificador o vector $\hat{\mathbf{s}}_i$ em sua substituição.

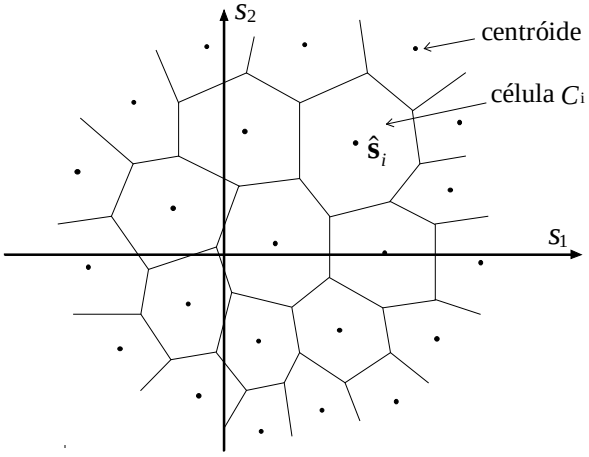


Figura 5-2: Ilustração da quantificação vectorial bidimensional.

A forma mais simples de QV pode ser encarada como uma generalização da quantificação escalar PCM, conhecida por VPCM (PCM vectorial). Na quantificação VPCM a palavra de código óptima é obtida através duma procura exaustiva, vector a vector, ao longo de todo o dicionário. Neste tipo de quantificador o número médio de *bits* por amostra é dado por

$$B = \frac{\log_2 L}{N}, \tag{5-5}$$

e a relação-sinal-ruído para o VPCM é dada por [Gersho (83)]

$$\mathrm{SNR}_N = 6B + K_N \quad \text{(dB)}. \tag{5-6}$$

Embora esta equação seja bastante similar à equação (5-3), a quantificação VPCM consegue uma maior SNR através do parâmetro K_N. A razão desta melhoria de desempenho prende-se com a possibilidade de este último codificador explorar a correlação entre vectores. Quando referentes à codificação da fala, os resultados obtidos em [Gersho (83)] revelaram que K_2 era maior do que K_1 em mais de 3 dB, enquanto que K_8 superava K_1 em mais de 8 dB.

Embora seja possível aumentar o ganho de codificação — aumento da taxa de compressão mantendo a qualidade de sinal, ou vice-versa — com o aumento de N e de

L, a complexidade computacional bem como as necessidades de memória crescem exponencialmente. Assim, o aumento do ganho de codificação à custa da QV apenas é viável até um determinado nível de complexidade, que dependerá sempre do tipo de aplicação. Normalmente, os benefícios da QV fazem-se sentir a taxas de 1 ou menos bits por amostra.

Para se projectar um dicionário, devem-se considerar essencialmente os seguintes aspectos: robustez do dicionário, eficiência do processo de procura, e escolha da medida de distorção.

O processo de povoamento do dicionário pode ser fixo ou adaptativo. Nos dicionários fixos as palavras de código são definidas *a prior* através dum processo que passa pela atribuição de um valor inicial a todos as palavras do dicionário, seguida por sucessivas correcções iterativas, usando para o efeito um grande número de vectores de treino. Normalmente, devem ser utilizados no mínimo 10 — preferencialmente 50 — vectores de treino por cada palavra de código.

A complexidade da QV de elevada dimensionalidade pode ser reduzida significativamente com o uso de dicionários estruturados que permitam uma eficiente procura, na maior parte dos casos em prejuízo do desempenho. Por exemplo, um método para construir um dicionário altamente estruturado consiste em formar as palavras de código por combinações lineares de uma pequena base de vectores. A complexidade pode igualmente ser reduzida normalizando os vectores do dicionário e codificando o ganho separadamente.

Uma vez que a fala resulta num sinal não estacionário, é desejável poder-se adaptar o dicionário de modo a acompanhar as suas propriedades estatísticas. Existem basicamente dois tipos de QV com dicionários adaptativos. A QV adaptativa em que a actualização do dicionário é baseada em amostras anteriores que também estarão disponíveis no descodificador; e a QV adaptativa em que o dicionário é actualizado com base em amostras correntes ou futuras, e como tal, deverá ser codificada informação adicional.

Para concluir, refira-se que os avanços conseguidos na QV estruturada de elevada dimensionalidade tem sido uma das principais razões para o acelerado progresso na codificação a baixos ritmos. Dicionários estocásticos, adaptativos e altamente estruturados são usados para codificar a excitação nos codificadores

híbridos, que são tidos como os principais responsáveis pela codificação de boa qualidade a baixos ritmos de transmissão.

5.2.2 Codificadores de Transformada e de Sub-banda

Nos algoritmos já referidos todo o processamento de codificação era realizado no domínio do tempo. Existem contudo codificadores, onde as redundâncias do sinal são exploradas no domínio das transformadas. São exemplo disso o codificador de sub-banda e o codificador de transformada. O que os distingue é a maneira como são obtidas as representações do sinal no domínio da frequência: enquanto que os codificadores sub-banda usam um banco de filtros, os codificadores de transformada servem-se de transformadas para sinais discretos. A possibilidade de redução do ritmo de transmissão encontra-se, em ambos os codificadores, na estrutura do espectro de potência localizada — espectro de potência aplicada, não à totalidade do sinal, mas apenas a um determinado segmento —, bem como nas qualidades perceptuais do ouvido humano.

❍ **Codificadores de sub-banda**

Nos codificadores sub-banda o espectro do sinal é dividido em várias sub-bandas, utilizando-se para o efeito um banco de filtros do tipo passa-banda, Figura 5-3. De seguida, a saída de cada filtro é sub-amostrada (decimada) e codificada separadamente. No descodificador o processo inverte-se: após desmultiplexagem do sinal, procede-se à descodificação da informação referente a cada uma das bandas seguida pela desmudelação. O processo de reconstrução do sinal é concluído após somadas todas as contribuições.

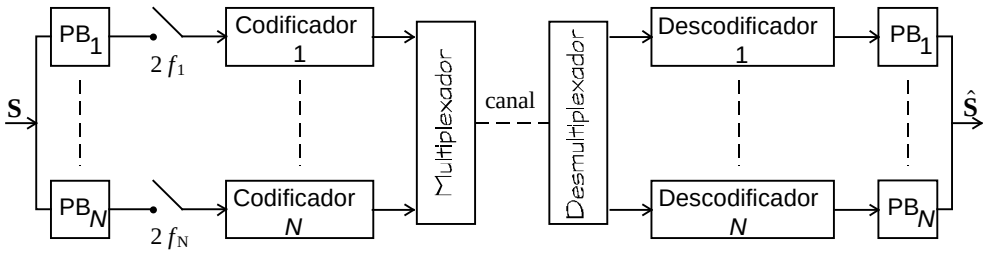


Figura 5-3: Codificador sub-banda típico.

Com este método, além do ruído inerente à quantificação, há ainda a considerar a distorção do tipo *aliasing* introduzida pelo processo de amostragem/desmudelação devido à sobreposição entre sub-bandas. Este codificador explora as propriedades estatísticas do sinal e ou critérios perceptuais, de modo a codificar o sinal usando diferentes quantidades de bits para cada sub-banda. Por exemplo, em sinais de voz atribui-se normalmente um maior número de bits às bandas de menores frequências, de modo a preservar informações importantes, como é o caso da estrutura do *pitch* e das formantes. Costuma-se igualmente utilizar, com o mesmo objectivo, sub-bandas mais estreitas nas baixas frequências. O desenho do banco de filtros é uma parte importante no projecto de um codificador deste tipo.

Este método de codificação deu já origem a dois standards: o standard AT&T *voice store-and-forward* usado para armazenamento de voz a 16 e 24 Kbps; e o standard G.722 a 64 Kbps da CCITT para audio a 7 KHz utilizado para teleconferência em RDIS. Enquanto que o primeiro standard utiliza um banco de filtros estruturado para 5 sub-bandas não uniformes, o segundo é baseado em apenas 2 sub-bandas.

❍ **Codificadores de transformada**

Na codificação por transformada, Figura 5-4, as componentes resultantes da aplicação duma dada transformada unitária localizada são quantificadas. Na descodificação o processo inverte-se: as componentes da transformada são obtidos por descodificação e em seguida utilizadas na reconstrução do sinal por intermédio da transformada inversa.

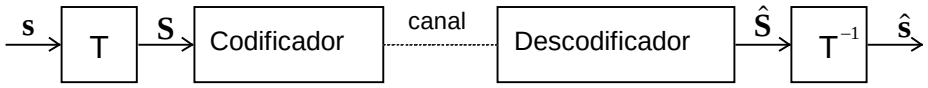


Figura 5-4: Codificador de transformada.

O potencial para a redução do ritmo de transmissão prende-se com a constatação de que as transformadas unitárias tendem a gerar componentes vizinhas não correlacionadas que podem, por isso, ser codificadas independentemente. Além disso, as componentes costumam variar lentamente com o tempo, podendo assim ser exploradas para extracção das redundâncias.

Neste codificador, cada segmento de sinal é convertido na sua transformada. Esta operação pode ser representada por uma multiplicação matricial, do tipo

$$\mathbf{S} = \mathbf{T}\,\mathbf{s}\,, \tag{5-7}$$

onde **s** e **S** representam vectores coluna contendo amostras, respectivamente do sinal de entrada e da sua transformada; e **T** é a matriz de transformação $N\times N$ que representa uma transformada unitária discreta, apropriada. A respectiva transformação inversa é então dada por

$$\mathbf{s} = \mathbf{T}^{-1}\,\mathbf{S}\,. \tag{5-8}$$

Esta equação é conhecida por equação de síntese e a anterior por equação de análise. Como **T** representa uma transformação unitária, a sua inversa será igual à matriz hermitiana, isto é, $\mathbf{T}^{-1} = \mathbf{T}^{H}$.

Os vectores coluna de $\mathbf{T}^{H}$ são entendidos como representando uma base de vectores, pois a partir da equação de síntese podemos admitir que o vector sinal resulta de uma combinação linear desses vectores. É claro que na ausência de ruído de quantificação e de erros de canal, a equação de síntese permite uma reconstrução exacta do sinal. Mas, como é sabido, pelo menos o ruído de quantificação está sempre presente, o que faz com que a precisão da reconstrução dependa, não só da variância do ruído, mas também do tipo de transformada.

Existem várias transformadas discretas que podem ser utilizadas, designadamente: a Transformada Discreta de Fourier (ou DFT), Transformada Discreta de Coseno (DCT), Transformada Walsh-Hadamard (WHT) e a Transformada Karhunen-Loéve (KLT). A KLT é óptima no que refere às componentes da transformada, devido a resultarem totalmente descorrelacionadas em qualquer tipo de sinal. No entanto, sendo a base de vectores da KLT formada pelos vectores próprios normalizados da matriz de autocorrelação do sinal, em muitos casos a sua implementação torna-se impraticável devido ao grande número de operações requeridas no cálculo desses vectores. Pelo contrário, as transformadas DFT e DCT, estando associadas com uma base de vectores sinusoidal, podem ser eficientemente implementadas usando a FFT. Também, sendo a matriz da WHT formada apenas por uns e menos uns, é possível obter-se um algoritmo eficiente para operar a transformação. O desempenho destas três transformadas fica um pouco aquém do obtido pela KLT. No entanto a DCT, tal como a DFT, tem também a vantagem de o

seu espectro expor as estruturas do *pitch* e das formantes. Por outro lado, com a DCT conseguem-se obter desempenhos próximos dos obtidos pela KLT, e para segmentos suficientemente grandes o desempenho da DFT aproxima-se do conseguido pela DCT e mesmo da KLT.

Utilizando estes métodos, alguns autores conseguiram ganhos consideráveis em relação ao PCM, nomeadamente obtiveram ganhos na ordem dos 9–10 dB para o KLT, 5 dB para o DFT, e 3 dB para o WHT. Um codificador que concilia o método DCT com a quantificação adaptativa usada na codificação das componentes da transformada, a funcionar entre 16–32 Kbps, foi proposto por [Zelinski (77)]. Com este codificador de transformada adaptativo (ATC) conseguiu-se um ganho de 17–23 dB acima do PCM logarítmico, e um ganho de 6 dB quando comparado com o ADPCM a 16 Kbps.

5.3 Codificadores Sinusoidais

Foram já apresentados algoritmos onde a codificação, ou era baseada directamente no sinal, ou então realizada a partir duma sua representação no domínio das transformadas. Quando o processo de codificação se baseia em representações sinusoidais da onda temporal ficamos na presença de uma outra classe de codificadores a que chamaremos codificadores sinusoidais. São disso exemplo os Codificadores de Transformada Sinusoidal (STC) e os codificadores de Excitação Multibanda (codificadores MBE). Embora as técnicas de compactação destes codificadores se baseiem em propriedades da voz — sendo por isso codificadores específicos da fala, tal como os *vocoders* —, tendem a ser mais robustos do que os tradicionais *vocoders* de dois estados de excitação, vozeado/não-vozeado, pois mantêm um razoável desempenho para uma grande variedade de sinais.

5.3.1 Análise-Síntese por Transformada de Fourier Localizada

A aplicabilidade das transformadas na análise-síntese da fala só faz sentido porque o sinal, sendo quase estacionário, pode ser modelado pelo seu espectro localizado. Uma vez que as suas características variam com o tempo, embora de uma forma lenta, o sinal de voz não pode ser representado pela transformada de Fourier

clássica; antes, deve-se utilizar uma transformação localizada para que possa acompanhar a evolução do sinal ao longo do tempo. A análise espectral variável pode então ser desempenhada pela Transformada de Fourier Localizada (TFL), dada por

$$S(n,\Omega) = \sum_{m=-\infty}^{+\infty} s(m)h(n-m)e^{-j\Omega m} = h(n) * s(n)e^{-j\Omega n} \tag{5-9}$$

onde $\Omega = wT = 2\pi fT$ é a frequência normalizada em radianos, e $h(n)$ é a janela de análise. A respectiva equação de síntese é dada por

$$s(m)h(n-m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(n,\Omega)e^{j\Omega m}d\Omega \tag{5-10}$$

que representa a TFL inversa.

O tamanho e forma da janela de análise controlam as resoluções temporal e espectral da TFL. Para a fala, o tamanho da janela situa-se normalmente entre os 5 e 20 ms, sacrificando-se dessa forma a resolução espectral da TFL.

Se calcularmos a TFL apenas para valores discretos de frequências, $\Omega_k = \Delta\Omega k\{k = 0,1,\cdots,N-1\}$, a expressão (5-9) converte-se na seguinte

$$S(n,\Omega_k) = \sum_{m=-\infty}^{+\infty} s(m)h(n-m)e^{-j\Omega_k m} = h(n) * s(n)e^{-j\Omega_k n}. \tag{5-11}$$

A partir da expressão (5-10) concluímos que se fizermos $m=n$ e $h(0)=1$ podemos obter uma aproximação de $s(n)$ através da seguinte equação de síntese

$$\tilde{s}_{TFL}(n) = \sum_{k=0}^{N-1} S(n,\Omega_k)e^{j\Omega_k n}. \tag{5-12}$$

Da equação (5-11) retira-se que cada componente da TFL pode ser encarada como sendo o resultado da excitação de um filtro linear, de resposta impulsional $h(n)$, pelo sinal de entrada sujeito a um determinado deslocamento na frequência. Isto permite-nos interpretar a TFL em termos de um banco de filtros, possibilitando dessa forma a existência de vários canais que podem ser codificados independentemente. Também, da expressão (5-12) deduzimos que a reconstrução do sinal é implementada somando as componentes de todos os canais depois de afectadas por um determinado avanço no tempo. Assim, o esquema de análise-síntese referente ao k ésimo canal, incluindo codificação e descodificação, encontra-se representado na Figura 5-5.

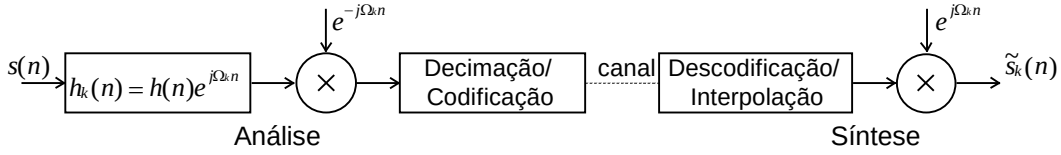


Figura 5-5: *k* ésimo canal do banco de filtros para a TFL.

O codificador "*vocoder* de fase" proposto por [Flanagan (66)] consistiu numa das primeiras tentativas de representação da fala explicitamente em termos da amplitude e fase do seu espectro localizado. Este codificador foi simulado com 30 canais, cobrindo uniformemente a banda de 50 a 3050 Hz. No actual *vocoder*, a derivada da fase é codificada, e no descodificador a fase é obtida por integração. Os codificadores usando métodos de análise-síntese de fala baseados na TFL têm produzido bons resultados, especialmente acima dos 14 Kbps.

5.3.2 Codificador de Transformada Sinusoidal

Num codificador de transformada sinusoidal típico, a fala é representada por uma combinação linear de *L* sinusóides de amplitudes, frequências e fases variáveis no tempo, isto é

$$\tilde{s}(n) = \sum_{k=1}^{L} A_k \cos(\Omega_k n + \phi_k) . \qquad (5\text{-}13)$$

O número de sinusóides, *L*, pode também variar de segmento para segmento. Depende normalmente do *pitch*.

A oportunidade de compactação associada a este tipo de codificador reside no facto de um sinal de voz, quando vozeado, ser altamente harmónico, podendo por isso ser representado por uma série apropriada de sinusóides; e quando não vozeado, a estrutura do seu espectro localizado pode ser preservada também por um modelo sinusoidal, com fases aleatórias definidas apropriadamente.

McAulay e Quatieri [McAulay (86)] deram uma grande contribuição para o desenvolvimento dos modelos sinusoidais. Mostraram ser possível reconstruir fala de grande qualidade usando sinusóides com amplitudes, frequências e fases correspondentes aos picos da TFL, realizada com janelas *Hamming* de largura 2.5 vezes maior que o *pitch* médio. Adicionalmente, são usados normalmente algoritmos

de interpolação para representação das amplitudes e fases de segmento para segmento. Tem-se verificado ainda que a síntese pode ser realizada eficientemente usando apenas oito sinusóides e a FFT com 1024 pontos actualizada ao fim de cada 10 ms.

Na codificação para baixos ritmos de transmissão, as frequências das sinusóides podem ser restringidas a inteiros múltiplos da frequência fundamental (*pitch*). Representando a frequência fundamental por Ω_0, a expressão (5-13) converte-se na seguinte

$$\tilde{s}(n) = \sum_{k=1}^{L(\Omega 0)} A_k \cos(k\Omega_0 n + \phi_k) \,. \tag{5-14}$$

A representação harmónica fornece uma série de frequências ideal apenas para segmentos perfeitamente vozeados. No entanto, mesmo para segmentos não vozeados, desde que as frequências se encontrem suficientemente próximas, a densidade espectral localizada do sinal pode ser igualmente preservada por uma série de frequências equidistantes. A opção pela representação através frequências equidistantes, além de permitir uma eficiente codificação das frequências das sinusóides, tem também como vantagem os parâmetros sinusoidais poderem ser convenientemente estimados por amostragem de uma DFT com alta resolução.

Na Figura 5-6 encontra-se representado um sistema básico de análise-síntese sinusoidal.

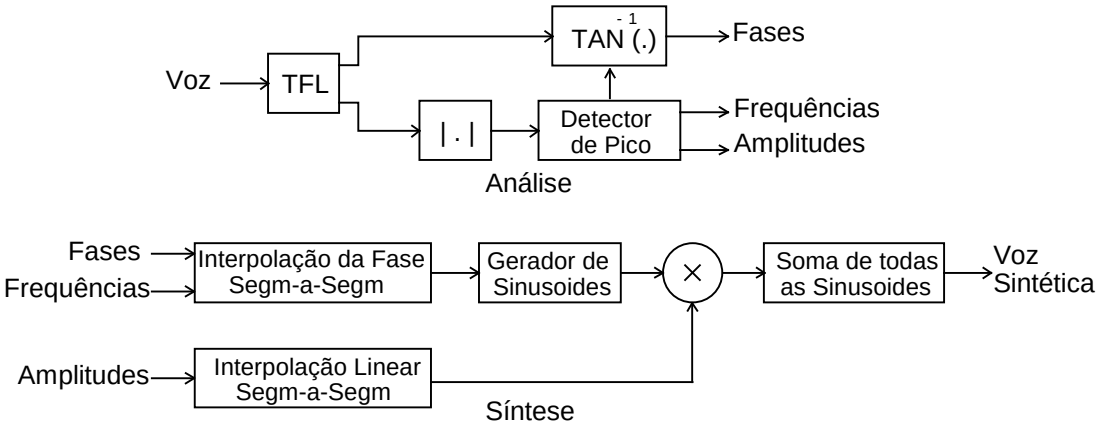


Figura 5-6: Análise-síntese de um sistema sinusoidal.

Os modelos sinusoidais têm sido utilizados com sucesso na codificação da voz a baixos ritmos. A análise-síntese do sistema sinusoidal tem um bom comportamento,

tanto em sinais de voz com ruído de fundo, como numa grande variedade de sinais, designadamente: múltiplos oradores, musica e sons biológicos.

Convém ainda referir que, embora os modelos apresentados tenham só por si uma importante utilidade na codificação, têm-se desenvolvido mais recentemente novas representações paramétricas robustas baseadas no sistema sinusoidal com o propósito de diminuir a sensibilidade ao ruído de quantificação e a erros de canal, principalmente quando se trata de codificação a baixos ritmos e de alta qualidade.

Por fim, saliente-se a contribuição que deram para este tipo de codificação Almeida e Silva [Almeida (84)] ao proporem um sistema de compressão dependente do *pitch*, e ainda o trabalho realizado por Marques, Almeida e Tribolet [Marques (90)] no desenvolvimento de um codificador harmónico a 4.8 Kb/s.

5.3.3 Codificador de Excitação Multibanda

Um codificador de Excitação Multibanda (MBE) baseia-se num modelo que decompõe o espectro localizado de um sinal de voz no produto do espectro de uma excitação pela envolvente espectral representativa do tracto vocal. Mais concretamente, o espectro dum sinal de voz é aproximado pelo seguinte produto

$$\hat{S}(\Omega) = H(\Omega)\left|X(\Omega)\right|, \tag{5-15}$$

onde $\left|X(\Omega)\right|$ representa a amplitude do espectro da excitação, e $H(\Omega)$ traduz a influência do tracto vocal, isto é, representa a envolvente do espectro localizado do sinal de voz, podendo ser obtida de uma forma aproximada por interpolação linear entre as amostras das harmónicas do espectro. Embora este tipo de representação espectral esteja implícito nos *vocoders* tradicionais de dois estados, a diferença reside no tipo de excitação. Nos codificadores MBE o espectro da excitação é formado por uma combinação de contribuições, quer harmónicas, quer de origem aleatória. Sendo este modelo de excitação formado ao longo da frequência pela concatenação de sub-bandas harmónicas e sub-bandas de natureza aleatória, está-se a supor que o vozeamento é dependente da frequência. Estas considerações fundamentam-se no facto de que o espectro dos sons mistos — fricativos vozeados, por exemplo — contem ambas as regiões, vozeadas e não vozeadas.

Portanto, no processo de análise o espectro do sinal de voz original é dividido em sub-bandas, e cada uma delas é declarada vozeada ou não vozeada. O número de sub-bandas é muito superior às utilizadas nos codificadores de sub-banda tradicionais, e pode ser escolhido de forma a ser igual ao número de harmónicas do espectro. Assim, o *pitch*, os parâmetros da envolvente e a informação de vozeamento para cada sub-banda, constituem a informação a extrair durante a codificação de um sinal de voz. Após a estimação do *pitch*, a excitação e a envolvente são estimadas simultaneamente por minimização do erro quadrático médio (LMS) entre o espectro original, $S(\Omega)$, e o sintético, $\hat{S}(\Omega)$, usando-se para o efeito um processo análise-por-síntese.

No processo de síntese as porções vozeadas da voz são sintetizadas no domínio do tempo usando um banco de sinusóides harmónicas, e as porções não vozeadas são determinadas aplicando a FFT a um segmente de ruído branco modelado por uma determinada "janela de análise", e por fim as amostras resultantes são multiplicadas pela envolvente espectral.

Baseados nesta técnica, desenvolveram-se codificadores MBE a 8 e a 4.8 Kbps [Hardwick (88)], tendo sido este último um dos candidatos para o standard DOD FS1016. A qualidade do codificador MBE de 4.8 Kbps foi avaliada em 92.7 e 60.4 com base nos critérios DRT e DAM respectivamente, e a sua complexidade foi estimada em 7 MIPS.

Mais recentemente foi proposta uma versão melhorada do codificador MBE (codificador IMBE) [Hardwick (91)] que emprega métodos mais eficientes na quantificação dos parâmetros do modelo MBE. Uma versão multi-ritmos (de 8, 4.8 e 2.4 Kbps) deste tipo de codificador foi implementada sobre um processador de sinal AT&T DSP32C, conseguindo-se codificação em tempo real. Desenvolveu-se também um IMBE a operar a 6.4 Kbps, que passou a fazer parte de standards para comunicação com os satélites AUSSAT e Inmarsat-M. O IMBE de 6.4 Kbps foi implementado numa DSP AT&T DSP32C com um atraso de 78.75 ms, e a sua classificação subjectiva revelou uma qualidade MOS de 3.4.

5.4 Codificadores de Fonte (*Vocoders*)

Contrariamente ao que se passa com os codificadores de forma de onda, já abordados, os codificadores que trataremos nesta secção são específicos da fala, e por isso, o seu desempenho degrada-se consideravelmente quando utilizados em quaisquer outros sinais. Uma vez que nestes codificadores a técnica de reconstrução do sinal de voz baseia-se geralmente num modelo representativo do mecanismo humano de produção de voz, os parâmetros extraídos de cada segmento de voz traduzem as características do modelo necessárias a reproduzir um segmento de amostras o mais semelhante possível com o segmento considerado.

Muito dos *vocoders* utilizam como excitação do sistema representativo do tracto vocal a excitação típica de dois estados, constituída por ruído aleatório e impulsos aproximadamente periódicos. Embora este tipo de excitação algo simplista esteja associado a ritmos de transmissão extremamente atractivos, é também em geral responsável por voz de qualidade sintética. É para contornar este facto que têm sido desenvolvidos modelos de excitação mais sofisticados, obtendo-se maior qualidade à custa do aumento da complexidade. Tanto a envolvente espectral como a excitação estão ambas relacionadas com a qualidade perceptual da voz sintetizada, porém, é a determinação dos parâmetros de excitação que mais contribui para esse efeito. Na estimação de ambos os parâmetros utilizam-se normalmente técnicas de estimação baseadas na predição linear ou em processamento homomorfo, ou então, a representação espectral do tracto vocal pode igualmente ser obtida por intermédio de um banco de filtros. São sobretudo estas distintas técnicas de estimação que diferenciam os *vocoders*, que passaremos a descrever. Refira-se apenas que a maior parte dos *vocoders* e codificadores híbridos fazem uso extensivo da predição linear; daí darmos especial ênfase aos conceitos associados a essa técnica.

5.4.1 *Vocoder* de Canal e *Vocoder* de Formante

Um *vocoder* de canal, em semelhança com o que acontece com muitos outros codificadores, representa o espectro da fala através do produto dos espectros da excitação e do tracto vocal. A representação da envolvente do espectro do tracto vocal é no entanto obtida usando um banco de filtros passa-banda. O número de canais

situa-se tipicamente entre 16 e 19, e a representação espectral torna-se tanto mais exacta quanto maior for o número de canais. A largura de banda dos canais é normalmente projectada de forma a aumentar com a frequência. Enquanto que a estrutura fina do espectro vozeado é representada usando como excitação uma série de impulsos espaçados pelo *pitch*, as zonas não vozeadas são produzidas usando ruído como excitação.

O *vocoder* de canal tem sido, no entanto, sujeito a vários melhoramentos. Em particular, a optimização dos *vocoders* de canal de baixos ritmos foi realizada através do aumento do número de canais, do uso de técnicas de alisamento espectral sobre o sinal de excitação, e explorando a correlação dos sinais de canal nos domínios do tempo e da frequência, usando técnicas do tipo DPCM.

O *vocoder* JSRU (*Joint Speech Research Unit*) [Holmes (80)] consiste numa das versões mais bem conseguidas do *vocoder* de canal. Este *vocoder* utiliza um banco de 19 filtros e serve-se de técnicas DPCM eficientes na codificação dos canais no domínio da frequência, com o objectivo de obter ritmos de transmissão de 2.4 Kbps. A qualidade do *vocoder* JSRU de 2.4 Kbps foi avaliada em cerca de 87 com base no critério DRT, e quando sujeito a 5% de ruído de transmissão o mesmo critério revelou uma qualidade de 81.

A principal diferença entre os *vocoders* de canal e os de formante é que as características ressonantes do banco de filtros dos *vocoders* de formante adaptam-se às trajectórias das formantes. As actuais implementações dos *vocoders* de formante utilizam configurações ressonantes, quer em paralelo, quer em cascata. Na Figura 5-7 encontra-se representado um modelo típico do *vocoder* de formante com configuração em cascata.

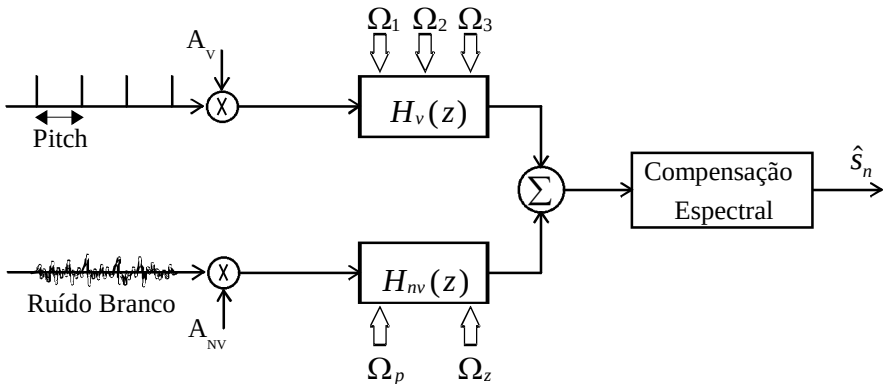


Figura 5-7: *Vocoder* de formante típico.

Este codificador inclui uma função de transferência $H_v(z)$ para a síntese da fala vozeada e uma segunda função de transferência, $H_{nv}(z)$, para a síntese da fala não vozeada. A função de transferência $H_v(z)$ consiste em três — L em geral — ressonâncias de segunda ordem do tipo AR, dispostas em cascata. Isto é

$$H_v(z) = \prod_{i=1}^{L} H_i(z), \tag{5-16}$$

com

$$H_i(z) = \frac{1 - 2e^{-\Omega_B(i)}\cos(\Omega_i) + e^{-\Omega_B(i)}}{1 - 2e^{-\Omega_B(i)}\cos(\Omega_i)z^{-1} + e^{-\Omega_B(i)}z^{-2}} \tag{5-17}$$

e onde Ω_i e $\Omega_B(i)$ denotam respectivamente a frequência e a largura de banda da *i* ésima formante. Para a fala não vozeada, $H_{nv}(z)$ consiste numa ressonância de segunda ordem AR (com o polo em Ω_p) em série com uma anti-ressonância AM também de segunda ordem (com o zero em Ω_z). A compensação espectral fixa, colocada à saída, modela os efeitos do impulso glotal e da radiação pelos lábios.

A dificuldade de implementação deste *vocoder* relaciona-se sobretudo com a estimação das formantes e das suas larguras de banda. Embora não se esteja actualmente a dar grande utilização aos *vocoders* de canal e de formante, toda a investigação já despendida nestes codificadores continua a dar ainda hoje valiosos contributos no desenvolvimento de novos *vocoders*.

5.4.2 *Vocoder* Homomorfo

A informação relacionada com a excitação bem como a que caracteriza o tracto vocal podem ser extraídas de um sinal de voz utilizando métodos de processamento homomorfo, como é o caso da desconvolução homomorfa. A ideia subjacente aos *vocoders* homomorfos prende-se com a constatação de que o logaritmo da amplitude do espectro da fala resulta numa combinação aditiva da amplitude logarítmica do espectro da excitação com a do tracto vocal, viabilizando dessa forma a separação das duas componentes da fala. Mesmo assim a sua separação ainda não é imediata uma vez que as amplitudes logarítmicas do espectro das componentes excitação e tracto vocal encontram-se sobrepostas. Felizmente, é possível mostrar que se aplicarmos a IFFT à amplitude logarítmica do espectro da fala, a sequência resultante passa a conter

as duas componentes separadas. Assim, um modelo genérico de análise-síntese que serve de base a este tipo de *vocoder* encontra-se esquematizado na Figura 5-8.

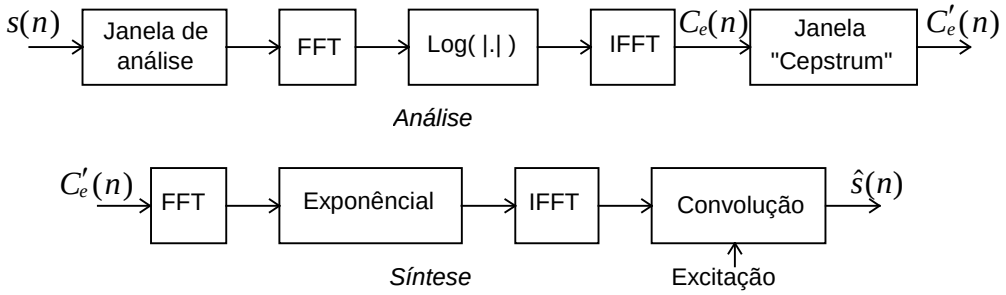


Figura 5-8: Sistema homomorfo de análise-síntese da fala.

O modelo apresentado serve-se da função *cepstral* — que engloba todo o processamento que acabámos de mencionar: FFT, Log(|.|) e IFFT — para desconvoluir as duas componentes da fala. As amostras da sequência *cepstral*, $C_e(n)$, que se encontram perto da origem — amostras $C'_e(n)$ — estão associadas com o tracto vocal e a informação referente à excitação evidencia-se sob a forma de um pico localizado a uma distância correspondente ao valor do *pitch*. Tal como se deduz da Figura 5-8, as amostras relacionadas com o tracto vocal podem ser isoladas usando uma janela *cepstral*, que deverá ter um comprimento inferior ao menor dos *pitchs* possíveis. Pode ainda ser mostrado que a sequência *cepstral* possui amostras de elevada amplitude localizadas em torno do *pitch* quando a fala é vozeada, sendo por isso possível estimar a frequência fundamental a partir da função *cepstral*.

Na síntese o processo inverte-se. Após a aplicação da FFT à sequência *cepstral* que contem apenas informação relativa ao tracto vocal, usa-se a função exponencial para converter a amplitude logarítmica em linear. De seguida aplica-se a IFFT ao espectro resultante de forma a se obter a resposta impulsional do tracto vocal, que é finalmente convoluida com o sinal de excitação para produzir a fala sintética.

Embora este *vocoder* não tenha encontrado muitas aplicações, alguns métodos de estimação do *pitch* e do tracto vocal baseados na função *cepstral* têm tido bastante aceitação noutras aplicações de processamento de voz. Adicionalmente, investigações sobre este tipo de técnica [Chung (89)] revelaram ser possível obter fala de boa qualidade a 4.8 Kbps, combinando a desconvolução homomorfa com um modelo de excitação do tipo análise-por-síntese.

5.4.3 *Vocoder* de Predição Linear

Os *vocoders* de Predição Linear (LP) representam a técnica de codificação mais investigada durante as duas últimas décadas. Nesta secção descreveremos os conceitos associados a este tipo de codificação. Apresentaremos ainda três possíveis modelos de excitação utilizados na Codificação por Predição Linear (LPC) em *loop* aberto.

Na Figura 5-9 encontra-se representado o sistema linear de produção de fala, desenvolvido por Fant [Fant (60)] com base no mecanismo de produção humano.

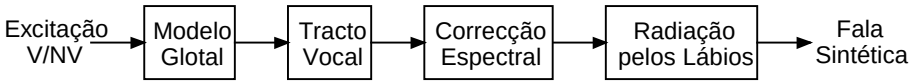


Figura 5-9: Modelo de produção de voz.

Para além da função de transferência referente ao tracto vocal — constituída apenas por pólos — inclui também o modelo glotal e o modelo de radiação labial: o primeiro implementado através de um filtro passa-baixo, e o segundo representado por $L(z) = 1 - z^{-1}$. Inclui ainda um factor de correcção espectral para compensação dos efeitos provocados pelos pólos mais elevados nas baixas frequências. Porém, a representação da fala nos vários métodos de codificação actuais, para além de omitir a correcção espectral, o zero da função de radiação labial é essencialmente cancelado por um dos pólos glotais, ficando assim o modelo reduzido a um sistema do tipo AR, ilustrado na Figura 5-10, e idêntico ao que foi deduzido no capítulo 4.

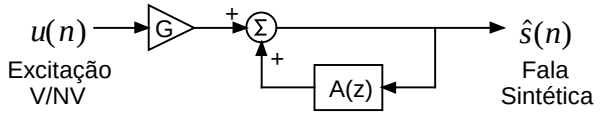


Figura 5-10: Modelo AR do sistema de produção de fala.

No domínio Z, este sistema é representado por

$$\hat{S}(z) = \frac{G}{1 - A(z)} U(z), \tag{5-18}$$

sendo a função de transferência constituída pelo ganho G e pelo polinómio de predição linear $A(z)$. Tanto os parâmetros da função de transferência como os referentes à excitação, não sendo conhecidos, devem ser determinados a partir de um

conjunto finito de amostras do sinal de voz. Os coeficientes de $A(z)$ são obtidos usando técnicas de predição linear (LP) [Makhoul (75)].

Um dos aspectos mais importantes no método LPC relaciona-se com a quantificação dos parâmetros LP. Em geral é suficiente considerar de 8 a 14 coeficientes LP. A quantificação directa destes parâmetros é normalmente evitada uma vez que o erro de quantificação pode conduzir à instabilidade do filtro de síntese. Embora a estabilidade esteja garantida quando quantificados os zeros de $(1 - A(z))$, estes também não representam uma boa opção. Além de o seu cálculo representar um esforço computacional bastante elevado, não formam uma série ordenada de parâmetros, não facilitando assim o desenvolvimento de modelos estatísticos para uma eficiente quantificação dos mesmos. Os coeficientes de reflexão, que podem ser obtidos a partir dos parâmetros LP por um processo recursivo [Rabiner (79)], pelo contrário, constituem uma série ordenada de parâmetros, e como já referido na secção 4.4, quando situados no intervalo de $]-1; 1[$ a estabilidade fica garantida. Por obedecerem a uma determinada ordem, isso é normalmente explorado codificando os primeiros coeficientes com mais precisão do que os últimos. De modo a se obter uma quantificação ainda mais eficiente, os coeficientes de reflexão são sujeitos, por sua vez, a uma transformação que dará como resultado uma outra série de parâmetros, com a particularidade de apresentarem menor sensibilidade à quantificação. O "logaritmo de razão de áreas" tem sido a transformação mais utilizada para esse efeito. Assim são os parâmetros resultantes da transformação

$$LRA(k) = \log \frac{1 + r_k}{1 - r_k} \tag{5-19}$$

que são por fim quantificados (r_k representa o k ésimo coeficiente de reflexão). Embora menos utilizadas, a transformação LSP (*Line Spectrum Pairs*) [Crosmer (85)], bem como a função inversa do seno ($Si(k) = \arcsin(r_k)$) têm também sido usadas por alguns algoritmos LPC na representação dos parâmetros LP.

Nos algoritmos LPC o tamanho da janela de análise é tipicamente de 20 a 30 ms, e os parâmetros são normalmente actualizados em intervalos de 10 a 30 ms. A procura da máxima compactação possível do sinal, passa pela utilização de segmentos de síntese de grande duração, implicando grandes transições nos parâmetros LP. Por isso, normalmente os segmentos são subdivididos, e os parâmetros dos subsegmentos

resultantes são obtidos por interpolação linear dos parâmetros LP de segmentos adjacentes. Estudos já realizados mostraram que se conseguem melhoramentos na qualidade da fala resultante usando interpolação dos coeficientes de predição [Atal (89)] ou dos parâmetros *LRA*'s [Vary (88)].

A excitação ideal no processo de síntese dum LPC seria o próprio resíduo de predição — obtido invertendo o sistema da Figura 5-10 e excitando-o com o sinal de voz original —; mas, devido à necessidade de compressão, utilizam-se em sua substituição modelos de excitação que de alguma forma se lhe assemelhem. No LPC clássico a excitação é modelada por uma sequência de impulsos periódicos para a fala vozeada, e uma sequência de ruído aleatório para fala não vozeada. A metodologia LPC pode também ser combinada com uma excitação mista; ou então, tal como acontece no método RELP (Predição Linear com Excitação Residual), a excitação pode ser gerada a partir da informação referente apenas à banda base do espectro do resíduo de predição.

De seguida descrevemos mais detalhadamente cada um dos três modelos de excitação, fazendo referência a alguns dos algoritmos de codificação em *loop* aberto que lhe estão associados. Os modelos de excitação do tipo análise-por-síntese serão abordados no capítulo seguinte.

❍ Modelos de Excitação de Dois Estados

Na excitação clássica de dois estados, o ganho, o *pitch* e o parâmetro binário de vozeamento constituem a informação a ser transmitida para o descodificador. O ganho referente a cada segmento tipicamente é determinado de modo a que a energia do segmento de fala sintética coincida com a do segmento original. A informação de vozeamento pode ser fornecida pelo algoritmo de detecção de *pitch*, ou então, pode ser determinada pela análise da energia e pela contabilização do número de vezes que o sinal passa por zero. Isto porque, os segmentos não vozeados são por norma segmentos de baixa energia, e onde ocorre grande número de passagens por zero. Para a estimação da frequência fundamental (*pitch*) existe uma grande diversidade de técnicas. Um método bastante poderoso, e por isso o mais utilizado, baseia-se na detecção do pico da sequência de autocorrelação do sinal sujeito a uma transformação não linear [Rabiner (79)]. O objectivo desta transformação é remover os efeitos da

função de transferência do tracto vocal, de modo a que as harmónicas do sinal passem a ter relativamente a mesma amplitude — técnica de alisamento espectral.

Na década de 70 foram desenvolvidas várias implementações em tempo real de codificadores LP. Desenvolveu-se um LPC a 4.8 Kbps que revelou uma qualidade de 87.3 segundo o critério DRT, seguido de um outro a 3.6 Kbps com um DRT de 87.6, e ainda um codificador LP de formante a 600 bps. Um outro codificador desenvolvido posteriormente para funcionar a 2.4 Kbps conhecido por LPC-10 tornou-se no standard FS-1015 [FS-1015 (84)]. Os critérios DRT e DAM revelaram uma qualidade para o LPC-10 de 90 e 48, respectivamente. Um algoritmo melhorado do LPC-10, o LPC-10e [Campbell (86)], foi proposto mais tarde, e evidenciou uma qualidade DRT de 89.9.

❍ **Modelos de Excitação Mista**

O modelo clássico de dois estados aborda o problema da excitação de forma algo simplista. Além das transições de vozeamento no sinal não terem qualquer possibilidade de serem fielmente representadas por uma excitação de apenas dois estados, os inevitáveis erros de decisão de vozeamento têm consequências desastrosas na qualidade final da fala. Foi precisamente na tentativa de colmatar estas imperfeições que surgiram os modelos de excitação mista, onde a excitação resulta de uma combinação de permanentes contribuições de ambas as naturezas: ruído e impulsos periódicos. Na Figura 5-11 encontra-se um possível modelo de excitação mista.

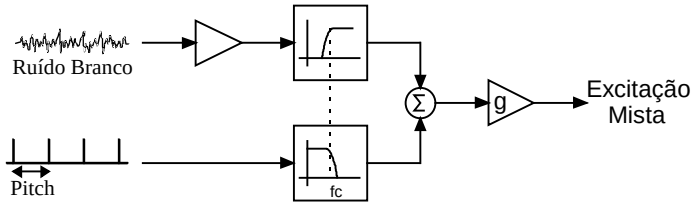


Figura 5-11: Modelo de excitação mista.

Neste modelo a sequência de impulsos excita a região de baixas frequências do filtro LPC de síntese enquanto que o ruído excita a zona de altas frequências. Os filtros passa-baixo e passa-alto e respectivos ganhos são escolhidos de forma a que a excitação adquira um espectro aproximadamente plano. A frequência de corte é a

mesma em ambos os filtros, e é estimada usando um detector de pico que determina a região do espectro associada a elevada periodicidade. É possível encontrarmos outros modelos de excitação mista mais elaborados, todavia baseiam-se essencialmente os mesmos princípios.

Um *vocoder* LPC de excitação mista a 2.4 Kbps, implementado numa DSP em tempo real, evidenciou uma qualidade perceptual de 58,9 valores segundo o critério DAM. Foi também implementado um *vocoder* a 4.8 Kbps [McCree (93)] a que correspondeu uma qualidade perceptual de 61.6, segundo o mesmo critério de qualidade.

❍ **Modelos de Excitação Residual**

Tal como ilustrado na Figura 5-12, o resíduo de predição, $e(n)$, traduz a excitação perfeita para a síntese LP. Este sinal transporta toda a informação não capturada pelo processo de análise LP, como é o caso da fase, informação do *pitch*, zeros devido aos sons nasais, etc.

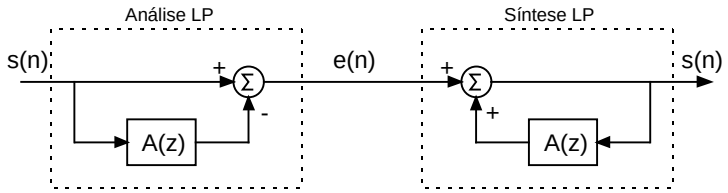


Figura 5-12: Análise-síntese LP usando o resíduo de predição.

Poder-se-á dizer que o processo de análise por predição linear comporta-se como um descorrelacionador de curta-duração, e por isso é de esperar que o resíduo de predição possua um espectro de potência relativamente plano. Existe uma classe de *vocoders* LP, conhecidos por codificadores RELP (Predição Linear com Excitação Residual), que se apoiam nesta constatação. Embora tanto os codificadores RELP como os codificadores ADPCM e APC se baseiam na codificação eficiente do resíduo de predição, a tecnologia RELP é distinta, pois a codificação do resíduo é baseada no espectro, e não na forma da onda temporal. A metodologia RELP fundamenta-se no princípio de que as componentes de baixa frequência da fala são perceptualmente as mais importantes. Assim um codificador RELP limita-se a enviar para o receptor apenas a banda base do resíduo de predição. Na Figura 5-13 encontra-se representado

um *vocoder* RELP a funcionar entre 6 e 9.6 Kbps, inicialmente proposto por [Magill (75)].

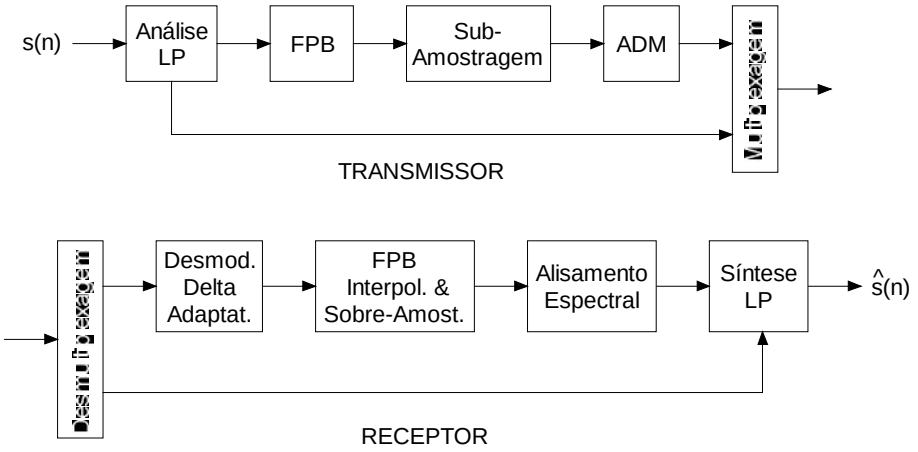


Figura 5-13: *Vocoder* RELP.

Este *vocoder* comprime a largura de banda do resíduo de predição a 800 Hz, possibilitando dessa forma a codificação da banda base do resíduo apenas a 5 Kbps. Depois de sujeito à filtragem passa-baixo, o resíduo é sub-amostrado e posteriormente codificado usando técnicas ADM. No receptor a banda base do resíduo é processada por uma transformação espectral não linear cuja função é regenerar as harmónicas de alta frequência. Por fim, a excitação do filtro de síntese é formada pela combinação do resíduo recuperado com uma quantidade adequada de ruído branco.

O resíduo de predição pode também ser codificado no domínio da frequência. Utilizando por exemplo a transformada FFT, um dos processos utilizados consiste em codificar e transmitir apenas as amplitudes e fases das componentes da FFT associadas às frequências da banda base.

O conceito da codificação do resíduo de predição tem ainda a seu favor o mascaramento auditivo conseguido. Com efeito, o espectro do ruído de quantificação do resíduo codificado ao ser moldado pelo filtro de síntese, fica automaticamente mascarado pela voz.

Verifica-se que, para taxas inferiores 4.8 Kbps, a qualidade conseguida por um codificador RELP é superior à obtida pelo codificador análogo com excitação de dois estados, devido essencialmente ao aproveitamento, por parte do primeiro, dos componentes do resíduo perceptualmente importantes. Ainda assim, a qualidade da fala num codificador RELP é limitada, principalmente pela perda de informação na

filtragem do resíduo para banda base. Os codificadores LP do tipo análise-por-síntese apresentados na secção seguinte evitam este problema utilizando modelos de excitação eficientes que podem ser optimizados segundo ambos os critérios: de aproximação dos sinais nas suas formas de onda, e em termos perceptuais.

5.5 Codificadores Híbridos

Os codificadores já referidos baseavam-se essencialmente numa de duas metodologias de codificação distintas. Facto esse que justificou a sua classificação em codificadores de forma de onda e codificadores paramétricos. Pelo contrário, os codificadores que vamos passar a descrever, designados por "codificadores híbridos", também conhecidos por "codificadores da terceira geração", combinam características de ambas as técnicas de codificação. Através do aproveitamento dos aspectos mais positivos de ambas as técnicas, consegue-se com esta nova metodologia aumentar o desempenho da codificação, tornando-se por isso na técnica de codificação actualmente mais adoptada. Estes codificadores, em analogia com os paramétricos, incluem modelos que traduzem de alguma forma a função do mecanismo humano de produção de voz, isto é, modelos responsáveis pela representação das estruturas formantes e harmónicas, típicas da fala. Têm ainda a particularidade de a sequência de excitação ser determinada por um processo de optimização em *loop* fechado, ou seja, segundo um processo do tipo análise-por-síntese. Este processo de optimização determina uma sequência de excitação que minimiza uma medida de erro representando a diferença entre o sinal de voz original e o sintetizado, normalmente afectada por uma "função perceptual". Sendo esta função projectada para realçar a informação perceptualmente importante, pretende-se com isso que a codificação seja optimizada para o ouvido humano. Este processo de optimização, para além de facultar um mecanismo que explora o sistema humano de audição, permite, através da minimização da diferença entre o sinal codificado e o original, uma aproximação entre os sinais, não apenas no que se refere às suas características espectrais, mas também em relação às suas formas de onda no tempo. Daí serem também reconhecidas a este tipo de codificação características de um "codificador de onda". Na Figura 5-14 encontra-se representado um codificador típico do tipo análise-por-síntese.

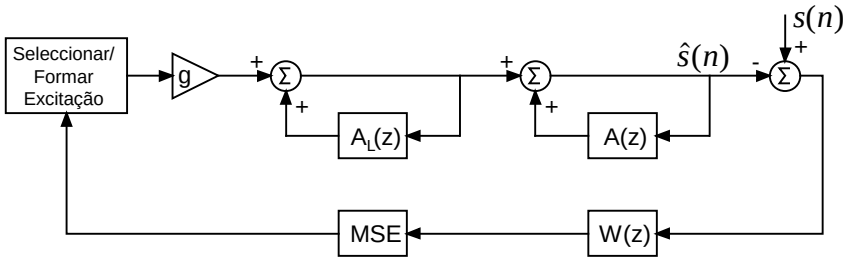


Figura 5-14: Codificador LP típico do tipo análise-por-síntese.

Para além do filtro de síntese LP de curta-duração, $H(z) = 1/\left(1 - A(z)\right)$, responsável pela introdução da estrutura formante da fala, o esquema inclui também um filtro de síntese LP de longa-duração, $H_L(z) = 1/\left(1 - A_L(z)\right)$, responsável pela introdução da estrutura harmónica relacionada com o *pitch*. Fazem igualmente parte do sistema um filtro perceptual, $W(z)$, que molda o erro de forma a que o ruído de quantificação fique parcialmente mascarado pelas formantes de elevada energia do sinal, e um gerador de excitação que gera ou selecciona uma sequência de excitação que minimize o erro quadrático médio (MSE - *Mean Squared Error*). A periodicidade de actualização do filtro LP de curta-duração situa-se tipicamente entre os 10 e os 30 ms, enquanto que no filtro LP de longa-duração o intervalo de actualização é cerca de metade. Embora o esquema análise-por-síntese mostrado na Figura 5-14 seja o mais usual, existem algumas configurações em *loop* fechado com a ordem dos filtros LP invertida, ou até mesmo desprovidas do filtro de predição de longa-duração. Note-se no entanto que a inclusão deste filtro na configuração em *loop* fechado contribuiu decisivamente para o grande aumento da qualidade conseguida neste tipo de codificação.

Passamos a descrever de seguida os algoritmos associados aos codificadores híbridos baseados nos três modelos de excitação mais comuns para a codificação análise-por-síntese: modelo multi-pulso, excitação por impulsos regulares, e modelo de excitação vectorial ou codificada.

5.5.1 Codificador Multi-Pulso

❍ **Codificador Original**

O algoritmo representado no diagrama da Figura 5-15 ilustra o processo de análise do codificador multi-pulso, proposto inicialmente por Atal e Remde [Atal (82)].

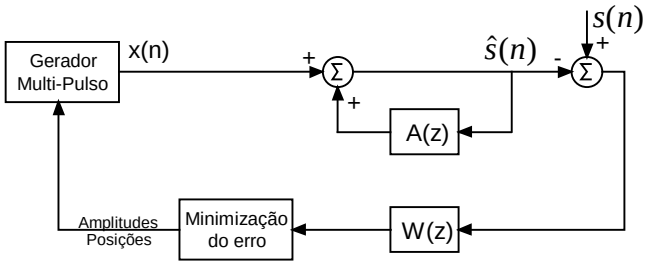


Figura 5-15: Codificador multi-pulso original.

Embora não possua o filtro LP de longa-duração, o que o caracteriza de facto é o seu modelo de excitação. Este gera uma sequência de impulsos com amplitudes e localizações determinadas sequencialmente, a um ritmo de 4 a 6 impulsos por cada 5 ms. Cada impulso é seleccionado separadamente por minimização do erro quadrático médio da distorção perceptual em relação a um segmento do sinal de entrada. Após determinado um impulso, a sua contribuição é de seguida subtraída ao sinal original, para que novo impulso possa ser calculado. A distorção perceptual pode ser representada por

$$e(n) = \left((s(n) - \hat{s}(n)\right) * w(n), \tag{5-20}$$

onde $w(n)$ representa a resposta impulsional referente à função de transferência do filtro perceptual, dada por

$$W(z) = \frac{1 - H(z)}{1 - H(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} \gamma^i a_i z^{-i}} \tag{5-21}$$

Este filtro tem como objectivo mascarar o ruído do sinal sintetizado de forma a torná-lo menos audível ou então menos desagradável ao ouvido humano. Sendo a densidade espectral do ruído do sinal sintetizado aproximadamente uniforme, verifica-se uma

elevada relação-sinal-ruído (SNR) nas regiões formantes, mas uma péssima entre formantes, onde a potência do sinal é baixa. Torna-se por isso preferível aumentar o ruído nas regiões formantes para um nível onde permaneça pouco perceptível, e à custa deste aumento, reduzir o ruído nas zonas entre formantes. É portanto com esse objectivo que se utiliza o filtro perceptual. No entanto, na escolha deste filtro particular teve-se também como preocupação possibilitar a diminuição da complexidade do processo de procura. Pois, com esta configuração verifica-se que se em vez de se calcular o erro instantâneo e só de seguida o pesar perceptualmente, comparar-se directamente o sinal original pesado com o sinal sintetizado pesado, o processamento envolvido resulta simplificado.

O parâmetro γ é o responsável pela deênfase da energia do erro nas regiões formantes. Embora possa tomar valores no intervalo $0 \le \gamma \le 1$, utiliza-se habitualmente um valor típico próximo de 0.8. A sua influência na largura de banda dos pólos de $W(z)$ traduz-se da seguinte forma

$$\Delta f = -\frac{1}{\pi T}\ln(\gamma) \quad (Hz), \tag{5-22}$$

de onde se conclui que quanto menor for o γ, maior será a largura de banda dos pólos, e consequentemente menor importância será dada à energia do erro situada nas zonas das formantes. A título elucidativo, note-se apenas que para $\gamma = 1$, $W(z) = 1$, e assim, embora a SNR seja máxima, não existindo mascaramento auditivo a voz sintetizada resulta de fraca qualidade para o ouvido humano. Se pelo contrário fizermos $\gamma = 0$, então $W(z) = 1/H(z)$. Neste caso o que é comparado é o sinal de excitação com o resíduo de predição de curta-duração e não o sinal reproduzido com o sinal original. Como há uma elevada concentração de ruído nas zonas formantes, a SNR é extremamente baixa, resultando um sinal sintetizado extremamente ruidoso.

Refira-se por fim que o modelo de excitação do algoritmo multi-pulso tem associado uma elevada carga computacional devido essencialmente a serem codificadas, além das amplitudes, as localizações dos impulsos. No entanto esta nova técnica de abordar a codificação permite fala descodificada de melhor qualidade do que a conseguida pelos *vocoders* LP clássicos. Por exemplo, com esta nova metodologia consegue-se boa qualidade de voz mesmo a ritmos na ordem dos 10 Kbps.

❍ Inclusão do Predictor de Longa-Duração

Um dos problemas do codificador multi-pulso inicialmente proposto prende-se no entanto com a sua dificuldade em codificar sinais de estrutura harmónica bastante acentuada (voz feminina, por exemplo), pois para este tipo de sinais o seu desempenho degrada-se consideravelmente. Devido à necessidade de ultrapassar esta dificuldade foi proposto mais tarde um codificador multi-pulso [Singhal (84)] incluindo um filtro LP de longa-duração para representação da estrutura harmónica do espectro da fala, tal como ilustrado na Figura 5-16.

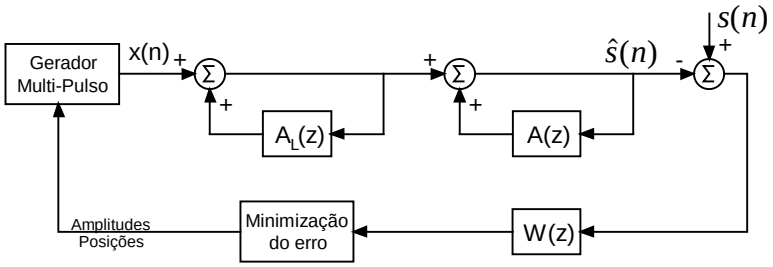


Figura 5-16: Codificador multi-pulso.

A inclusão deste novo filtro LP fundamenta-se na constatação de que o resíduo de predição de curta-duração — aquele que se obtém por filtragem do sinal original por $(1 - A(z))$ — contém ainda periodicidade semelhante à do sinal original, significando existir ainda uma considerável correlação entre amostras desfasadas por uma distância igual à do *pitch*. Assim a utilização de um filtro recursivo com predictor de longa-duração torna-se desejável para retirar esse tipo de correlação ainda existente no sinal. O comportamento do predictor de longa-duração traduz-se pela predição de cada uma das amostras a partir, não das amostras imediatamente anteriores, mas de amostras que ocorram um "período" mais cedo. Uma possível configuração para a respectiva equação de síntese deste filtro poderá ser dada por

$$x_L(n) = x(n) + a_L x_L(n - \tau). \tag{5-23}$$

Embora nesta equação apenas se utilize um coeficiente de predição, é também usual utilizarem-se 2 ou 3 coeficientes. Uma das vantagens da utilização de mais do que um coeficiente é tornar o ganho do predictor dependente da frequência, pois para altas frequências a correlação é menor do que para baixas. Para além disso, tem também como vantagem efectuar-se interpolação, uma vez que o período da fundamental não contém normalmente um número inteiro de intervalos de amostragem. Contudo, é

óbvio que a utilização de um maior número de coeficientes tem também como resultado o aumento do ritmo de transmissão.

Neste codificador, os parâmetros do predictor de longa-duração são determinados a partir do resíduo de predição de curta-duração obtido por filtragem do sinal original pelo filtro inverso de curta-duração, $(1-A(z))$. O *pitch*, τ, pode ser determinado através da detecção do pico da sequência de autocorrelação do resíduo de predição, e o ganho, a_L é determinado usando a expressão $a_L = \hat{r}_{ee}(\tau)/\hat{r}_{ee}(0)$, com $\hat{r}_{ee}()$ representando a função de autocorrelação do resíduo de predição.

Com a configuração incluindo os dois filtros recursivos LP conseguem-se acréscimos de 6 a 10 dB em termos de SNR. Por exemplo, a 10 Kbps consegue-se uma SNR de 17 dB.

5.5.2 Codificador de Impulsos Regulares (RPE)

Os codificadores RPE (*Regular Pulse Excitation*) utilizam também sequências de excitação compostas por múltiplos impulsos. No entanto, os impulsos nestes codificadores encontram-se uniformemente espaçados, e por isso as suas posições podem ser determinadas especificando apenas a localização do primeiro impulso dentro do segmento considerado e a distância entre dois impulsos consecutivos. Mas como o número de impulsos por segmento é normalmente fixo, o espaçamento entre impulsos não necessita de ser codificado. Utilizam-se tipicamente de 10 a 13 impulsos por segmento de 5 ms, e a localização do primeiro impulso é actualizada ao fim de cada 5 ms. As amplitudes dos impulsos são determinadas através da resolução de um sistema de equações lineares.

A optimização do tipo análise-por-síntese nos algoritmos RPE considera um esquema de filtragem inversa, onde o resíduo é obtido com a excitação do filtro de curta-duração inverso, $(1-A(z))$, pelo sinal de voz original, tal como ilustrado na Figura 5-17. O resíduo de curta-duração passa então a ser representado por uma dada sequência de impulsos regulares seleccionada por um processo de minimização do erro perceptual. O gerador de excitações produz *n* sequências desfasadas, de *N* amostras com *N/n* impulsos igualmente espaçados. A Figura 5-18 ilustra um exemplo

de 4 sequências de 5 ms com 10 impulsos cada, representando um possível conjunto de segmentos utilizado pelo gerador de excitação do RPE.

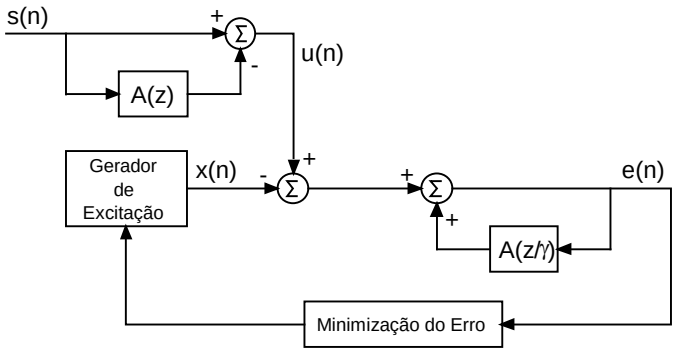


Figura 5-17: Análise RPE.

K=0 1...1...1...1...1...1...1...1...1...1...

K=1 .1...1...1...1...1...1...1...1...1..

K=2 ..1...1...1...1...1...1...1...1...1.

K=3 ...1...1...1...1...1...1...1...1...1

Figura 5-18: Possíveis sequências de excitação (N=40, n=4).

O processo de selecção implícito no diagrama da Figura 5-17 entra apenas em linha de conta com o filtro de predição de curta-duração. Contudo, incluindo um filtro de predição de longa-duração verifica-se também neste codificador um aumento considerável do seu desempenho, particularmente para vozes femininas. Por exemplo, um sofisticado esquema de codificação baseado na tecnologia RPE incluindo um predictor de longa-duração foi adoptado para o standard europeu para rádio móvel (GSM - *Group Special Mobile*) [Vary (88)].

5.5.3 Codificador CELP

Os codificadores RPE e multi-pulso já abordados conseguem reprodução de voz de boa qualidade a ritmos de transmissão médios. Para se conseguir a mesma qualidade a baixos ritmos de transmissão torna-se necessário uma representação mais eficiente das sequências de excitação. São vários os codificadores que se baseiam em técnicas do tipo análise-por-síntese e utilizam modelos de predição linear para extraírem as redundâncias do sinal a codificar, no entanto apenas o codificador CELP

(*Code Excited Linear Predictive*) acrescenta às técnicas referidas uma outra que consiste na quantificação vectorial gaussiana do sinal de excitação. Este codificador de excitação vectorial permite reproduzir fala codificada a baixo ritmo de transmissão com qualidade comparável à conseguida pelos codificadores de forma de onda a médios ritmos. Daí podermos afirmar ser este o método de codificação que transpôs a lacuna que existia entre os codificadores de forma de onda e os *vocoders*.

O modelo de excitação do CELP consiste num dicionário composto por um conjunto de vectores *N* dimensionais formados por amostras pseudo-aleatórias, pretendendo representar ruído branco com distribuição gaussiana. Daí este codificador ser também conhecido por codificador estocástico. O codificador inicialmente proposto por Schroeder e Atal [Schroeder (84)], ilustrado na Figura 5-19, utiliza um dicionário de 1024 vectores (palavras de código) de 5 ms de duração.

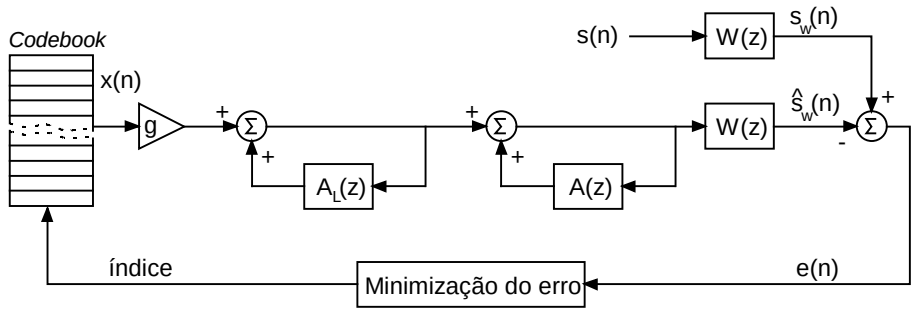


Figura 5-19: Codificador CELP.

O codificador é portanto constituído por dois filtros recursivos lineares, contendo um deles, na malha de realimentação, o predictor de longa-duração e o outro o predictor de curta-duração. A procura da palavra de código óptima consiste num procedimento do tipo análise-por-síntese, sendo por isso o algoritmo de codificação constituído por um bloco de síntese formado pelos dois filtros, ganho e excitação. O dicionário é percorrido exaustivamente, de forma a encontrar-se o vector de excitação que minimiza a distorção perceptual do sinal sintetizado. Durante este processo, cada palavra do dicionário, após ter sido afectada pelo ganho, *g*, é filtrada pelos dois filtros LP recursivos obtendo-se um sinal que será subtraído ao original. A diferença resultante é processada de forma a obter-se um erro quantitativo que será tido em conta na procura da melhor palavra. O processamento do erro consiste numa filtragem perceptual seguida do cálculo do erro quadrático médio. Essa filtragem, como

sabemos, é responsável pela deênfase da energia do erro nas regiões formantes, pois nestas regiões o ruído de quantificação é parcialmente mascarado pela voz. Um filtro normalmente utilizado que produz o efeito pretendido é dado por

$$W(z) = \frac{1 - \sum_{i=1}^{P} a_i \gamma_1^i z^{-i}}{1 - \sum_{i=1}^{P} a_i \gamma_2^i z^{-i}} \qquad com \quad 0 \le \gamma_2 \le \gamma_1 \le 1. \tag{5-24}$$

A função de transferência deste filtro perceptual é mais geral do que a utilizada nos codificadores análise-por-síntese convencionais — equação (5-21). Novamente os parâmetros γ_1 *e* γ_2 têm como função controlar a energia do erro nas regiões formantes. Têm-se conseguido bons resultados em termos de qualidade perceptual para $\gamma_1 = 0.9$ e $\gamma_2 = 0.6$.

Note-se que, de modo a simplificar o processo de procura, a filtragem perceptual é aplicada directamente sobre o sinal de entrada e sobre a fala sintetizada, e só depois calculada a diferença entre os sinais, tal como se depreende por observação do diagrama expresso na Figura 5-19. Assim, no processo de análise, o vector erro *N*×1 resultante do *k* ésimo potencial vector de excitação é dado por

$$\mathbf{e}(k) = \mathbf{s}_w - \hat{\mathbf{s}}_w^0 - \hat{\mathbf{s}}_w(k), \tag{5-25}$$

onde $\mathbf{s}_w$ é o vector *N*×1 que contém as amostras do sinal de voz filtrado perceptualmente; $\hat{\mathbf{s}}_w^0$ o vector que contém a saída do filtro cascata $H(z)H_L(z)$ devido ao seu estado inicial; e $\hat{\mathbf{s}}_w(k)$ o vector de voz sintética gerada pelo *k* ésimo vector do dicionário (*codebook*) e filtrada perceptualmente. A minimização do erro quadrático médio é obtida começando por minimizar $\varepsilon(k) = \mathbf{e}^T(k)\mathbf{e}(k)$ em relação ao ganho g_k. Obtendo-se assim o ganho

$$g_k = \frac{\left(\mathbf{s}_w - \hat{\mathbf{s}}_w^0\right)^T \hat{\mathbf{s}}_w(k)}{\hat{\mathbf{s}}_w^T(k)\hat{\mathbf{s}}_w(k)}. \tag{5-26}$$

Considerando este ganho, o erro quadrático médio passa finalmente a ser expresso por

$$\varepsilon(k) = \left(\mathbf{s}_w - \hat{\mathbf{s}}_w^0\right)^T \left(\mathbf{s}_w - \hat{\mathbf{s}}_w^0\right) - \frac{\left[\left(\mathbf{s}_w - \hat{\mathbf{s}}_w^0\right)^T \hat{\mathbf{s}}_w(k)\right]^2}{\hat{\mathbf{s}}_w^T(k)\hat{\mathbf{s}}_w(k)}. \tag{5-27}$$

Como o erro quadrático, $\varepsilon(k)$, é uma grandeza positiva, a sua minimização é conseguida maximizando apenas o 2º termo do 2º membro da equação, pois o 1º termo não depende do vector de excitação — termo constante. Assim, é seleccionado o k ésimo vector do dicionário, $\mathbf{x}(k)$, que maximize essa grandeza; e o respectivo ganho, g_k, pode por fim ser obtido voltando à equação (5-26).

Embora os parâmetros do predictor de longa-duração possam ser determinados usando uma configuração em *loop* aberto, a opção pelo método em *loop* fechado conduz a fala de melhor qualidade. Neste segundo processo os parâmetros do predictor de longa-duração são determinados antes dos parâmetros de excitação. É por exemplo o caso do *pitch*. Parâmetro cuja gama de variação se situa tipicamente entre os inteiros de 20 a 147. Pode-se, no entanto trabalhar com valores de *pitch* não inteiros, pois eficientes predictores de longa-duração com resolução subamostral (valores de *pitch* não inteiros) evidenciam desempenhos semelhantes ou superiores aos conseguidos por predictores de maior ordem, mas com valores inteiros de *pitch*.

De forma a reduzir ao máximo o ritmo de transmissão, têm-se desenvolvido inclusive técnicas de interpolação para este tipo de codificadores. A interpolação é utilizada não apenas nos parâmetros de predição de curta-duração, que é o mais habitual nestes codificadores, mas também nos parâmetros de predição de longa-duração, bem como nos parâmetros de excitação. A qualidade dos codificadores CELP pode ainda ser melhorada através de pós-processamento. Habitualmente à fala descodificada aplica-se um processo de filtragem de forma a realçar as estruturas harmónicas e formantes. Uma configuração típica deste tipo de processamento consiste numa cascata de filtros perceptuais de longa-duração e de curta-duração com ganhos e parâmetros de expansão de largura de banda (γ`s) apropriados.

Uma das principais desvantagens — senão a principal — do CELP original reside na sua complexidade, e consequentemente no esforço computacional requerido na procura da palavra de código do dicionário. Muitos dos algoritmos CELP requerem processadores capazes de executarem 20 MIPS e memória na ordem dos 40 KB. Embora estas exigências não representem hoje qualquer obstáculo significativo, ainda há bem pouco tempo eram factores bastante limitativos. Isso justificou um grande esforço no desenvolvimento de dicionários altamente estruturados de forma a permitir processos de procura eficientes.

Com a codificação CELP consegue-se qualidade de comunicação a 8 Kbps, e têm-se alcançado progressos significativos de forma a se conseguir boa qualidade a ritmos inferiores a 4.8 Kbps, e em particular a 4 Kbps para o standard Norte Americano para Telefones Celulares. Existem actualmente vários tipos de algoritmos CELP que fazem parte de standards para comunicações. É o caso, por exemplo, do standard FS1016 CELP a funcionar a 4.8 Kbps adoptado pelo Departamento de Defesa Norte Americano, do codificador VSELP (*Vector-Sum Excited Linear Prediction*) a 8 Kbps adoptado pelo Sistema Celular Digital Norte Americano, e o standard G.728 *Low-Delay* CELP a 16 Kbps desenvolvido pela CCITT.

Em jeito de conclusão, pode-se afirmar que a tecnologia CELP "quebrou" a barreira dos 9.6 Kbps que foi considerada ao longo de vários anos como o ritmo de transmissão mínimo para fala de qualidade de comunicação.

Capítulo 6

Modelos Variáveis no Tempo

No capítulo anterior fez-se um levantamento das várias metodologias de codificação convencionais com o intuito de nos munirmos dos conhecimentos necessários sobre o tema em que se enquadra o presente trabalho. Não se pretende com este trabalho o desenvolvimento exaustivo de um novo algoritmo de codificação, mas sim elaborar um estudo sobre a utilização de modelos variáveis no tempo, visando a optimização da codificação. Com esse propósito analisaremos o desempenho de um vocoder de predição linear de parâmetros variáveis.

6.1 A Conveniência dos Modelos Variáveis

Como sabemos, o modelo digital de produção de fala anteriormente deduzido, que serve de base aos principais algoritmos de codificação de fala hoje adoptados, foi directamente obtido a partir de um modelo físico, representando o sistema humano de produção de voz. Para que o modelo assim obtido reproduza fielmente o sistema modelado, é evidente que as suas características terão que se alterar ao longo do tempo, de forma a traduzir correctamente a evolução que caracteriza o sistema humano durante a produção de fala. O modelo convencional simula esse comportamento do sistema humano; no entanto, fá-lo de uma forma aparentemente pouco correcta, uma vez que os parâmetros do sistema são adaptados, não de uma forma gradual e contínua, como acontece no sistema humano, mas sim através de actualizações periódicas. No fundo, embora se saiba que a fala é um sinal não estacionário, ao aplicar-se este tipo de modelo, está-se implicitamente a supor que o segmento de sinal em análise é estacionário. Esta aproximação baseia-se no facto das características do sinal variarem lentamente ao longo do tempo. Optando-se assim por dividi-lo numa sequência de segmentos relativamente pequenos, nos quais o sinal passa a ser considerado estacionário, de modo a se poder aplicar o modelo de

parâmetros constantes a cada um desses segmentos. Os parâmetros são portanto mantidos fixos ao longo de cada segmento de sinal em análise, o que não se encontra em conformidade com a natureza contínua da evolução do sistema humano. Assim, parece preferível utilizar um modelo de parâmetros variáveis no tempo, em vez do modelo convencional de parâmetros fixos.

Iremos portanto desenvolver o nosso estudo baseado-nos num modelo em que os coeficientes LP se alteram gradualmente, em vez de se alterarem descontinuamente em intervalos fixos, acompanhando dessa forma "amostra a amostra" a evolução do sinal a modelar. Ao conseguirmos por esta via obter um sistema que modele de uma forma mais coerente o processo natural da produção de fala, será então lícito esperar melhorias na relação qualidade/compactação dum codificador de voz baseado neste tipo de modelo, em relação a um outro que utilize o modelo com parâmetros fixos.

6.2 Modelação com base em Funções *B-spline*

Uma forma de implementar o modelo variável no tempo é assumir que cada um dos parâmetros do modelo seja representado por uma combinação linear de uma série de funções pré-definidas, variáveis no tempo [Louis (75)] [Hall (83)]. Assim, se usarmos q+1 funções, $\{ f_k(n) \}$, na representação de um parâmetro α, que se pretende variável ao longo de um segmento de N amostras, esse parâmetro será expresso por

$$\alpha(n) = \sum_{k=0}^{q} a_k f_k(n), \quad \text{com } n = 0, 1, ..., N\text{-}1. \tag{6-1}$$

Desta forma os parâmetros do modelo passam a ser também funções no tempo, contrariamente ao que se verificava no modelo fixo, onde os parâmetros se mantinham constantes ao longo de cada segmento em análise.

Com a escolha adequada das funções de base, pode-se aproximar com razoável precisão uma grande diversidade de variações associadas a uma variável. No entanto, se permitirmos que os parâmetros variem arbitrariamente, teremos os mesmos graus de liberdade no modelo paramétrico e no sinal original, não resultando desse modo qualquer compactação de dados. É importante por isso conhecermos a natureza das variações de um sinal de voz.

No projecto do modelo variável, a escolha das funções de base pode ser determinante em termos de resultados finais. Por razões óbvias, pretende-se que o sinal possa ser modelado a partir de um número de funções tão reduzido quanto possível. Podem ser utilizadas, entre outras, funções polinomiais e funções trigonométricas como é o caso das séries de Fourier [Hall (83)] [Amir (89)]. Existe contudo um conjunto de funções, conhecidas por funções *B-spline*[1], que embora tenham já dado bons resultados em diversas aplicações, como por exemplo, utilizadas como funções interpoladoras em processamento de imagem [Unser (91)] [Aldroubi (92)] [Ries (91)] [Engels (88)], as suas potencialidades ainda não foram devidamente exploradas na modelação paramétrica de um sinal de voz. Apesar destas funções serem de natureza polinomial, não têm os principais inconvenientes desse tipo de funções. Isto porque, sendo funções de suporte finito, dão origem a curvas segmentalmente polinomiais, permitindo dessa forma o uso de segmentos polinomiais de baixa ordem. Evita-se assim as limitações associadas aos polinómios de ordem mais elevada: grande peso computacional e probabilidade de ocorrência de instabilidades. Possuem ainda a vantagem, pelo menos em relação às séries de Fourier, de serem funções reais, e por isso de mais fácil tratamento.

É nosso objectivo estudar o desempenho de um codificador de voz tendo por base um modelo de parâmetros variáveis representados por combinações lineares de funções *B-spline*. Em apêndice encontram-se alguns fundamentos teóricos associados a este tipo de funções, bem como os relacionados com a formalização de curvas a partir dessas funções ou de outras da mesma família. É o caso das curvas do tipo *Spline* e das curvas *Bézier*, que tendo muito em comum com as curvas *B-spline*, o seu estudo é importante, não apenas para melhor entendimento da natureza das *B-splines*, mas sobretudo para realçar certas características importantes, específicas apenas dessas funções. Assim, todas as considerações e deduções que se seguem terão sempre como suporte o referido apêndice.

Note-se que, se fizermos representar a forma de onda de um sinal por uma curva bidimensional, onde o eixo das ordenadas traduza a amplitude do sinal e o eixo das abcissas a evolução no tempo, os princípios que regem a procura de uma forma de representar a evolução de um parâmetro variável no tempo são semelhantes aos

[1] *Basis Spline*

utilizados para descrição de curvas. Assim, existem essencialmente duas maneiras de abordar o problema da reconstrução de uma curva:

- através da técnica *Curve Fitting*: processo através do qual se cria uma curva, obrigando-a a passar por uma série de valores amostrados, sendo por isso normalmente utilizado para fazer interpolação;
- através da técnica *Curve Fairing*: processo utilizado quando necessário construir a curva sem conhecimento *a priori* de qualquer ponto por onde a curva deva passar.

Ambas as técnicas poderão ser exploradas no âmbito deste trabalho. Por isso, representam desde já uma forte razão para optarmos pelas funções *B-spline* em detrimento das outras, pois a formalização das curvas *B-spline* pode ser conseguida segundo as duas técnicas, o que não acontece com as restantes: as curvas *Spline* são obtidas apenas segundo a técnica *Curve Fitting*, e as *Bézier* segundo a técnica *Curve Fairing*. As curvas *Bézier* têm ainda como inconveniente a dificuldade que existe em ligar segmentos de curva entre si, pois de modo a garantir a continuidade[2] C^2 em toda a extensão da curva resultante é necessário utilizar segmentos polinomiais de ordem bastante elevada. Para além disso, as *B-splines* têm a vantagem de serem funções com suporte finito. Por isso, uma base de funções *B-spline* tem um comportamento não global, que significa que cada função *B-spline* exerce uma influência apenas local na curva reproduzida, não afectando assim a forma da curva em pontos distantes da zona de influência.

A teoria de funções *B-spline* foi inicialmente sugerida por Schoenberg e posteriormente utilizada por Gordon e Riesenfeld na definição de curva. A *k* ésima função *B-spline* normalizada de ordem *r*, $F_k^{(r)}(t)$, encontra-se definida pela formula recursiva de Cox-deBoor em apêndice na equação (A-36).

Para além dos aspectos já referidos, a principal propriedade associada às funções *B-spline* de ordem *r* é servirem de base ao subspaço de funções segmentalmente polinomiais de ordem *r* e contínuas em C^{r-2} em toda a sua extensão. Qualquer função $\varphi_r(t)$ pertencente ao subspaço referido pode ser expressa através da seguinte relação

[2] Diz-se que uma função tem continuidade C^r quando ela própria e todas as suas derivadas até à ordem *r* são funções contínuas.

$$\varphi_r(t) = \sum_{k=-\infty}^{+\infty} a_k F_k^{(r)}(t). \tag{6-2}$$

A função $\varphi_r(t)$ é univocamente determinada pelos seus coeficientes *B-spline* $\{a_k\}$. É precisamente a suavidade destas funções e o suporte finito das funções de base que tornam as *B-splines* atractivas na modelação paramétrica de sinais não estacionários.

Note-se que todas as *B-splines* com a mesma ordem *r* têm a mesma forma, sendo cada uma delas apenas uma versão deslocada de qualquer uma outra — ver Figura A-11 em apêndice. Portanto, qualquer *B-spline* de ordem *r* é obtida deslocando a função $F_0^{(r)}(t)$, ou seja,

$$F_k^{(r)}(t) = F_0^{(r)}(t-k)\,. \tag{6-3}$$

A abordagem feita em apêndice, refere-se apenas a funções *B-spline* contínuas. Interessa-nos, no entanto, definir as *B-splines* discretas, obtidas directamente através da amostragem das respectivas funções contínuas.

A função *B-spline* normalizada contínua de ordem *r*, $F_0^{(r)}(t)$, encontra-se definida em apêndice na equação (A-41). Como pretendemos convertê-la numa função discreta, torna-se necessário expandir a função por um factor *m*, para que possamos fazer a amostragem a uma frequência unitária. O produto da ordem, *r*, pelo factor de expansão horizontal, *m*, representa o número de amostras onde a função *B-spline* se encontra definida e, como será mostrado mais adiante, o factor de expansão é dado pela seguinte relação

$$m = \frac{N}{(q+1)-r+1}, \tag{6-4}$$

onde *N*, em analogia com a equação (6-1), é o número de amostras do segmento considerado, (*q*+1) é o número de funções de base a utilizar na reconstrução do parâmetro, e *r* é a ordem das funções de base *B-spline*.

Fazendo $t = t_0 + nT$, e para simplificar se considerarmos $t_0 = 0$ e $T = 1$, obtém-se a *B-spline* discreta de ordem *r*, expandida horizontalmente por um factor *m*,

$$b_0^{(r)}(n) \stackrel{\Delta}{=} F_0^{(r)}\left(\frac{t_0 + nT}{m}\right) = F_0^{(r)}\left(\frac{n}{m}\right) \Leftrightarrow$$

$$b_0^{(r)}(n) = \frac{1}{m^{r-1}} \sum_{j=0}^{r} \frac{(-1)^j}{(r-1)!} \binom{r}{j} (n - jm)^{r-1} \mu(n - jm) \tag{6-5}$$

Similarmente ao caso contínuo, é também possível obter uma *B-spline* discreta expandida por um factor *m*, de qualquer ordem, através duma sequência de convoluções [Unser (91)],

$$b_0^{(r)}(n) = \frac{1}{m^{r-1}} \overbrace{b_0^{(1)}(n) * b_0^{(1)}(n) * \cdots * b_0^{(1)}(n)}^{r\times} * F_0^{(r)}(n) \,. \tag{6-6}$$

Comparando esta expressão com a expressão análoga (A-39) referente ao caso contínuo, verificamos que existe nesta uma convolução adicional com a *B-spline* discreta não expandida de ordem *r*. Esta convolução adicional serve para garantir que os valores da *B-spline* discreta coincidam com a *B-spline* contínua nos respectivos pontos de amostragem.

Depois de termos definido explicitamente a função *B-spline* discreta $b_0^{(r)}(n)$ em (6-5), e sabendo que todas as outras da mesma ordem *r* se obtêm deslocando esta ao longo do eixo das abcissas com *shifts* múltiplos de *m*, podemos finalmente expressar o parâmetro variável que pretendemos modelar como uma combinação linear dessas funções. Assim, em analogia com o caso contínuo expresso na equação (6-2), se α representar o parâmetro a modelar com *B-splines*, será expresso da seguinte forma

$$\alpha(n) = \sum_{k=-\infty}^{+\infty} a_k b_k^{(r)}(n) \,, \tag{6-7a}$$

com $\quad b_k^{(r)}(n) = b_0^{(r)}(n - km) \,. \tag{6-7b}$

Como as funções $b_k^{(r)}(n)$ têm suporte temporal finito, o somatório (6-7a) pode ser transformado num somatório finito. Assim, uma vez que $b_0^{(r)}(n) = 0$ fora do intervalo $0 \le n < mr$, a função $b_0^{(r)}(n - km)$ encontra-se definida apenas no seguinte intervalo[3]

$$0 \le n - km < mr \quad \Leftrightarrow \quad floor\left(\frac{n}{m}\right) - r + 1 \le k \le floor\left(\frac{n}{m}\right). \tag{6-8}$$

Como estamos a considerar um segmento de sinal de tamanho *N*, o parâmetro α será constituído por *N* amostras, e portanto podemos considerar $n = 0,1,\ldots,N-1$. Tendo em conta a gama de variação do *n* em conjugação com a desigualdade (6-8) resulta o seguinte intervalo de variação para o parâmetro *k*,

[3] *floor*(*x*) representa o maior inteiro menor ou igual a *x*.

$$floor\left(\frac{0}{m}\right) - r + 1 \le k \le floor\left(\frac{N-1}{m}\right) \tag{6-9}$$

Para optimizarmos o número de funções de base[4] em relação ao tamanho do segmento de síntese, podemos impor que *N* seja múltiplo de *m*, obtendo desta forma os limites do somatório (6-7a),

$$1 - r \le k \le floor\left(\frac{N}{m} - \frac{1}{m}\right) \Leftrightarrow 1 - r \le k \le \frac{N}{m} - 1, \tag{6-10}$$

vindo portanto

$$\alpha(n) = \sum_{k=1-r}^{N/m-1} a_k b_k^{(r)}(n). \tag{6-11}$$

Repare-se ainda que podemos ajustar o incrementador *k*, de modo a ficarmos com o somatório na forma prevista em (6-1),

$$\alpha(n) = \sum_{k=0}^{N/m+r-2} a_{k-r+1} b_{k-r+1}^{(r)}(n). \tag{6-12}$$

Ficamos então na posse do modelo pretendido em (6-1) com

$$f_k(n) = b_{k-r+1}^{(r)}(n) = b_0^{(r)}\left(n - (k+1-r)m\right), \text{ e} \tag{6-13}$$

$$q = \frac{N}{m} + r - 2. \tag{6-14}$$

Esta última relação dá-nos precisamente o factor de expansão *m* apresentado anteriormente na expressão (6-4).

No cálculo de cada amostra *n* de um dado parâmetro α, os limites do somatório (6-12) poderão ainda ser alterados para

$$0 \le n - (k+1-r)m < mr \quad \Leftrightarrow \quad \frac{n}{m} - 1 < k \le \frac{n}{m} - 1 + r \quad \Leftrightarrow$$

$$floor\left(\frac{n}{m}\right) \le k \le floor\left(\frac{n}{m}\right) - 1 + r. \tag{6-15}$$

Portanto, para calcularmos a amostra *n* do parâmetro α, necessitamos apenas dos coeficientes a_k com o índice *k* contido no intervalo dado em (6-15).

Embora um modelo variável no tempo envolva um maior número de parâmetros que o modelo convencional — vários coeficientes a_k do modelo variável por cada parâmetro α do modelo fixo —, permite dividir o sinal em segmentos de análise

[4] O número de funções de base é dado pelo intervalo de variação do parâmetro *k*.

bastante mais longos, sendo por isso possível com o novo método reduzir o número total de parâmetros necessários para codificar a totalidade do sinal de voz.

Só será possível escolher o número de *B-splines* à posteriori numa fase experimental, testando o desempenho do modelo para diferentes quantidades de funções de base. O mesmo se pode dizer em relação à escolha da ordem das *B-splines*. No entanto, tendo em conta o parecer de alguns autores, embora referentes a outro tipo de aplicações [Silva (93)] [Unser (91)], é de esperar que a ordem r=3 seja suficiente.

6.3 Modelação dos Coeficientes LP

Durante a produção da fala o tracto vocal altera-se de uma forma contínua, umas vezes lentamente, outras mais rapidamente. Como a função do filtro LP (filtro de Predição Linear) é essencialmente modelar o comportamento do tracto vocal, devido à natureza contínua da evolução deste será de esperar melhor desempenho por parte de um modelo baseado num filtro LP de coeficientes variáveis do que o obtido por um modelo de coeficientes fixos.

O modelo LP de parâmetros fixos foi deduzido no capítulo 4, e encontra-se representado no diagrama ilustrado na Figura 4-15. Este sistema tem a seguinte representação no tempo

$$\hat{s}(n) = \sum_{i=1}^{p} c_i \hat{s}(n-i) + Gu(n). \tag{6-16}$$

O somatório desta expressão corresponde a uma estimativa da amostra $s(n)$ do sinal de voz. Por isso, este sistema é conhecido por modelo de Predição Linear, pois cada uma das amostras do sinal de voz é predita através de uma combinação linear das p amostras imediatamente anteriores. Isto fundamenta-se na constatação de que num sinal de voz existe grande correlação entre amostras vizinhas. O erro cometido nesta predição é em parte corrigido adicionando uma outra parcela dada pelo produto do ganho G com a amostra $u(n)$ do sinal de excitação.

Como estamos interessados, não neste modelo, mas sim num de parâmetros variáveis, é necessário converter cada um dos parâmetros fixos c_i em parâmetros variáveis $c_i(n)$. Assim, para o modelo variável, a equação (6-16) transforma-se na seguinte

$$\hat{s}(n) = \sum_{i=1}^{p} c_i(n)\hat{s}(n-i) + Gu(n). \qquad (6\text{-}17)$$

Com o propósito de tornar um parâmetro variável no tempo, na secção anterior descreveu-se um processo de exprimir um dado parâmetro α a partir de uma combinação linear de funções *B-spline*. Assim, se na equação (6-1) o parâmetro α representar um qualquer coeficiente c_i, este passará a ser expresso por

$$c_i(n) = \sum_{k=0}^{q} c_{ik}\, f_k(n), \quad (\text{para } i=1,..,p), \qquad (6\text{-}18)$$

em que $f_k(n)$ representa a k ésima função de base (função *B-spline*) dada pelas equações (6-13) e (6-5), e $q+1$ é o número de funções utilizadas, dado pela equação (6-14). Os coeficientes constantes c_{ik}'s são os parâmetros que pretendemos extrair do sinal de voz, onde o índice i é uma referência ao coeficiente variável $c_i(n)$ e o índice k é uma referência à função de base $f_k(n)$.

Ao assumirmos que os coeficientes do filtro LP são combinações lineares de (q+1) funções pré-definidas { $f_k(n)$, k=0,1,..,q}, estamos a manter a linearidade do problema, ou seja, a estimação dos parâmetros c_{ik}'s do modelo LP continua a passar pela resolução de um sistema de equações lineares. Esta característica é a principal vantagem deste tipo de modelos [Louis (75)] [Hall (83)] [Silva (93)] [Grenier (91)].

Como já referido anteriormente, o uso de funções de base do tipo *B-spline* permite que a extracção dos seus coeficientes (coeficientes c_{ik}'s, neste caso) se possa processar segundo duas técnicas alternativas:

❍ **Técnica *Curve Fitting***

A modelação dos coeficientes LP pode simplesmente ser encarada como resultando da aplicação de interpolação — não necessariamente linear — aos coeficientes LP fixos do modelo LPC convencional. Isto é, a curva *B-spline* que descreverá a forma de variação de um dado parâmetro LP é determinada de modo a passar por todos os valores desse parâmetro, previamente determinados para vários subsegmentos de análise consecutivos — técnica *Curve Fitting*. Assim o processo de determinação dos coeficientes c_{ik}'s das *B-splines* passa, em primeiro lugar, pela

determinação dos coeficientes c_i's do modelo LPC fixo — parâmetros da equação (6-16) — para vários subsegmentos consecutivos.

Os parâmetros fixos c_i's, sendo determinados pelo processo convencional, obtêm-se através da minimização do erro quadrático médio (MSE) correspondente à diferença entre as amostras do sinal original e as amostras estimadas à custa das anteriores, por perdição linear, e sobre um intervalo finito (subsegmento de análise). Portanto, o erro é dado por

$$MSE = \sum_{n=0}^{L-1} \left(s(n) - \hat{s}(n) \right)^2 , \tag{6-19}$$

com

$$\hat{s}(n) = \sum_{i=1}^{p} c_i s(n-i), \tag{6-20}$$

em que L é o comprimento do subsegmento de análise, e $s(n)$ é o valor da n ésima amostra desse segmento. Assim, os coeficientes c_i's são obtidos para cada subsegmento a partir da minimização do erro dado pela equação (6-19). Como deduzido em [Rabiner (79)], chega-se ao seguinte sistema de equações

$$\sum_{k=1}^{p} c_k \sum_{n=0}^{L-1} s(n-k)s(n-i) = \sum_{n=0}^{L-1} s(n)s(n-i) , \quad 1 \le i \le p \tag{6-21}$$

Existem vários métodos eficientes para resolver este sistema, nomeadamente os métodos de autocorrelação e de covariância [Rabiner (79)]. O método de autocorrelação parece ser o mais aconselhável, essencialmente devido a termos a possibilidade de garantirmos com este método a estabilidade do filtro LP de síntese, e também devido à sua simplicidade de implementação. Refira-se contudo que, a estabilidade do filtro de síntese refere-se à implementação do modelo fixo — equação (6-16) —, pois em relação à sua versão de parâmetros variáveis — equação (6-17) — não é possível estabelecer qualquer garantia de estabilidade. Mesmo assim é importante garantirmos a estabilidade dos parâmetros fixos para que a probabilidade de ocorrências de instabilidades nos respectivos parâmetros variáveis seja menor.

Após encontrados os pontos por onde as curvas *B-splines* devem passar — coeficientes c_i's calculados para vários subsegmentos consecutivos — o procedimento a seguir consiste em aplicar a técnica *Curve Fitting* de forma a determinar os coeficientes *B-spline*, c_{ik}'s, que geram as respectivas curvas. Assim, se pretendermos

modelar os parâmetros LP ao longo de um segmento de N amostras ($s(0)$, $s(1)$, ...,$s(N)$), os coeficientes LP fixos terão que ser calculados para (q+1) subsegmentos de tamanho L centrados, respectivamente, em torno das amostras $s(m-rm/2)$, $s(2m-rm/2)$, ..., $s((q+1)m-rm/2)$. Como resultado obtemos q+1 conjuntos de coeficientes LP, que passarão a ser identificados por $\{c_i^{(j)}\}$, em que o coeficiente $c_i^{(j)}$ representa o coeficiente c_i calculado no j ésimo subsegmento de sinal. Como pretendemos que a curva que representa a forma do parâmetro variável $c_i(n)$ (com i=1, 2, ..,p) passe pelos coeficientes fixos $c_i^{(0)}$, $c_i^{(1)}$, ..., $c_i^{(q)}$, obviamente, em analogia com o exposto em apêndice, estamos a impor as seguintes igualdades: $c_i(m-rm/2)=c_i^{(0)}$, $c_i(2m-rm/2)=c_i^{(1)}$, ..., $c_i((q+1)m-rm/2)=c_i^{(q)}$. Substituindo o primeiro termo de cada uma destas igualdades pela expressão dada na equação (6-18), resulta o seguinte sistema de equações, análogo ao da equação (A-44),

$$\begin{array}{lcl} c_i^{(0)} & = & c_{i0}f_0(m-rm/2) \quad +\cdots \quad +c_{iq}f_q(m-rm/2) \\ c_i^{(1)} & = & c_{i0}f_0(2m-rm/2) \quad +\cdots \quad +c_{iq}f_q(2m-rm/2) \\ & \vdots & \\ c_i^{(q)} & = & c_{i0}f_0((q+1)m-rm/2) \quad +\cdots \quad +c_{iq}f_q((q+1)m-rm/2) \end{array} \qquad (6\text{-}22)$$

Resolvendo este sistema para cada um dos coeficientes LP (para i=1,2..p) encontramos finalmente os coeficientes *B-spline*, c_{ik}'s, necessários à formalização dos parâmetros variáveis. O i ésimo parâmetro variável virá então dado pela expressão da equação (6-18).

❍ Técnica *Curve Fairing*

Pretendemos agora determinar os coeficientes das *B-splines* sem qualquer conhecimento *a priori* de pontos por onde devam passar as respectivas curvas — técnica *Curve Fairing*. De modo a encontrarmos a expressão que traduza a obtenção dos coeficientes c_{ik}'s a partir de um sinal de voz $s(n)$, consideremos a seguinte notação vectorial

$$\begin{array}{l} \mathbf{S}(n)=\left[\mathbf{S}_0^T(n)\ldots\mathbf{S}_q^T(n)\right]^T, \quad \text{com } \mathbf{S}_k(n)=\left[s(n-1)f_k(n)\ldots s(n-p)f_k(n)\right]^T, \\ \text{e} \quad \mathbf{C}=\left[\mathbf{C}_0^T\ldots\mathbf{C}_q^T\right]^T, \quad \text{com } \mathbf{C}_k=\left[c_{1k}\ldots c_{pk}\right]^T. \end{array} \qquad (6\text{-}23)$$

Tal como no LPC de parâmetros fixos, o critério para obtenção dos coeficientes óptimos $\mathbf{C}^*$ consiste em minimizar o erro quadrático médio [Hall (83)] dado por

$$MSE = \sum_{n=0}^{N-1} \left[s(n) - \tilde{s}(n) \right]^2, \tag{6-24}$$

com

$$\tilde{s}(n) = \sum_{i=1}^{p} c_i(n) s(n-i), \tag{6-25}$$

onde N é o tamanho do segmento de sinal em estudo.

O somatório em (6-25) pode ainda ser transformado no produto interno entre os dois vectores $\mathbf{C}$ e $\mathbf{S}(n)$. Substituindo na expressão (6-25) o parâmetro $c_i(n)$ dado em (6-18), obtém-se

$$\begin{aligned} \tilde{s}(n) &= \sum_{i=1}^{p} \sum_{k=0}^{q} c_{ik} f_k(n) s(n-i) \\ &= \sum_{k=0}^{q} \sum_{i=1}^{p} c_{ik} f_k(n) s(n-i) = \sum_{k=0}^{q} \mathbf{C}_k^T \mathbf{S}_k(n) = \mathbf{C}^T \mathbf{S}(n) . \end{aligned} \tag{6-26}$$

Como demonstrado em [Haykin (91)], os coeficientes são obtidos resolvendo o seguinte sistema de equações normais

$$\mathbf{R}\mathbf{C}^* = \mathbf{P}, \tag{6-27}$$

onde $\mathbf{R}$ é a matriz de autocorrelação, dada por

$$\mathbf{R} = \sum_{n=0}^{N-1} \mathbf{S}(n) \mathbf{S}^T(n), \tag{6-28}$$

$\mathbf{P}$ é o vector de correlações–cruzadas, com

$$\mathbf{P} = \sum_{n=0}^{N-1} s(n) \mathbf{S}(n) \tag{6-29}$$

e $\mathbf{C}^*$ é o vector de coeficientes óptimo, ou seja, o conjunto de parâmetros que desejamos obter a partir do sinal de voz. São estes os únicos parâmetros que necessitam de ser determinados, pois todas as funções de base são fixas e conhecidas.

Capítulo 7

Implementações e Resultados

No Capítulo anterior descrevemos, em traços gerais, como modelar os parâmetros LP a partir de funções *B-spline*. As duas técnicas mencionadas — *Curve Fitting* e *Curve Fairing* — têm características próprias que lhes conferem diferenças em qualidades de capital importância. Como referido, no método *Curve Fitting*, correspondendo a uma técnica de interpolação, os coeficientes *B-spline* — coeficientes c_{ik} da equação (6-18) — são determinados a partir dos coeficientes fixos do modelo LPC convencional. Por outro lado, a determinação dos coeficientes pelo método *Curve Fairing* processa-se directamente a partir do sinal de voz. Isto permite-nos, por si só, fazer algumas considerações importantes em relação às duas técnicas alternativas. Repare-se que, aquando da determinação dos coeficientes LP fixos, os seus valores são optimizados partindo-se do princípio que os coeficientes LP se vão manter fixos ao longo de cada subsegmento, não havendo por isso a garantia de que, ao se aplicar a técnica *Curve Fitting*, esses valores conduzam às curvas *B-spline* ideais. Pelo contrário, aplicando a técnica *Curve Fairing*, os coeficientes *B-spline* são obtidos directamente através da minimização do erro (6-24); o mesmo será dizer que os valores encontrados são aqueles que melhor aproximam — segundo o critério utilizado — o sinal reconstruído do sinal original. Daí podermos considerar que as curvas *B-spline* assim obtidas serão as ideais. Existe ainda outra forma de interpretarmos esta diferença: enquanto que no método *Curve Fairing* a análise LP é realizada com base em parâmetros descritos por curvas *B-splines*, no método *Curve Fitting* isso não acontece, pois as *B-splines* apenas são consideradas na síntese LP.

Por tudo o que foi dito, são por demais evidentes as vantagens inerentes à utilização do método *Curve Fairing* para o tipo de implementação por nós pretendida. Assim, todos os testes que realizaremos terão sempre por base implementações incluindo este tipo de técnica.

7.1 Determinação dos Parâmetros LP

A fase crucial na utilização das *B-splines* é exactamente a determinação dos coeficientes c_{ik} da expressão (6-18), de forma a que os valores dos parâmetros variáveis $c_i(n)$, que desejamos reproduzir, minimizem o erro dado pela equação (6-24).

Comecemos por construir a matriz de autocorrelação **R** definida em (6-28). Será então dada por

$$\mathbf{R} = \sum_{n=0}^{N-1} \mathbf{S}(n)\mathbf{S}^T(n) = \underbrace{\begin{bmatrix} \mathbf{R}_{00} & \mathbf{R}_{01} & \cdots & \mathbf{R}_{0q} \\ \mathbf{R}_{10} & \mathbf{R}_{11} & \cdots & \mathbf{R}_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{q0} & \mathbf{R}_{q1} & \cdots & \mathbf{R}_{qq} \end{bmatrix}}_{p(q+1)\times p(q+1)} \tag{7-1}$$

com a submatriz $\mathbf{R}_{kl}$ sendo dada por

$$\mathbf{R}_{kl} = \sum_{n=0}^{N-1} \mathbf{S}_k(n)\mathbf{S}_l^T(n) = \underbrace{\begin{bmatrix} \phi_{kl}(1,1) & \phi_{kl}(1,2) & \cdots & \phi_{kl}(1,p) \\ \phi_{kl}(2,1) & \phi_{kl}(2,2) & \cdots & \phi_{kl}(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{kl}(p,1) & \phi_{kl}(p,2) & \cdots & \phi_{kl}(p,p) \end{bmatrix}}_{p\times p} \tag{7-2}$$

onde cada elemento desta matriz corresponde a uma função de correlação dada por

$$\phi_{kl}(i,j) = \sum_{n=0}^{N-1} s(n-i)s(n-j)\, f_k(n)\, f_l(n), \tag{7-3}$$

com as variáveis *i* e *j* referindo-se a atrasos das amostras do sinal, e os índices *k* e *l* referenciando respectivamente as funções de base $f_k(n)$ e $f_l(n)$.

Por sua vez, o vector de correlações–cruzadas **P** definido em (6-29) pode ser representado por

$$\mathbf{P} = \sum_{n=0}^{N-1} s(n)\mathbf{S}(n) = \begin{bmatrix} \mathbf{P}_0^T & \mathbf{P}_1^T & \cdots & \mathbf{P}_q^T \end{bmatrix}^T, \tag{7-4}$$

onde cada um dos vectores $\mathbf{P}_k$ é dado por

$$\mathbf{P}_k = \sum_{n=0}^{N-1} s(n)\mathbf{S}_k(n) = \begin{bmatrix} \varphi_k(1) & \varphi_k(2) & \cdots & \varphi_k(p) \end{bmatrix}^T, \tag{7-5}$$

correspondendo cada um dos elementos a uma nova função de correlação dada por

$$\varphi_k(i) = \sum_{n=0}^{N-1} s(n)s(n-i)\, f_k(n), \tag{7-6}$$

com a variável *i* referindo-se ao desfasamento entre amostras, e o índice *k* à função de base $f_k(n)$.

Dependendo do tipo de segmento de análise utilizado, obtêm-se dois métodos de análise LP distintos [Rabiner (79)]. Um é o método de autocorrelação, no qual se assume que o sinal em análise é nulo fora do segmento de síntese $0 \le n \le N-1$. O outro é o método de covariância, onde, ao não se considerar a simplificação anterior, deve-se também entrar em linha de conta na análise LP com algumas amostras anteriores ao segmento de síntese. Enquanto que no primeiro, o segmento de análise coincide com o segmento de síntese, no segundo método o segmento de análise contém também p amostras anteriores às do segmento de síntese. Isto é, no método de covariância considera-se o segmento de análise $-p \le n \le N-1$. Como na implementação do método de autocorrelação se assume que o sinal é estacionário, o método de covariância sem *windowing* é o mais indicado [Hall (83)] para o tipo de aplicação pretendida.

O sistema de matrizes (6-27) para obtenção dos coeficientes óptimos pode então ser descrito por

$$\underbrace{\begin{bmatrix} \mathbf{R}_{00} & \mathbf{R}_{01} & \cdots & \mathbf{R}_{0q} \\ \mathbf{R}_{10} & \mathbf{R}_{11} & \cdots & \mathbf{R}_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{q0} & \mathbf{R}_{q1} & \cdots & \mathbf{R}_{qq} \end{bmatrix}}_{p(q+1)\times p(q+1)} \begin{bmatrix} \mathbf{C}_0^* \\ \mathbf{C}_1^* \\ \vdots \\ \mathbf{C}_q^* \end{bmatrix} = \begin{bmatrix} \mathbf{P}_0 \\ \mathbf{P}_1 \\ \vdots \\ \mathbf{P}_q \end{bmatrix} \tag{7-7}$$

Estamos na presença de um sistema de $p(q+1)$ equações e $p(q+1)$ incógnitas, sendo portanto de prever um sistema de grandes dimensões e uma consequente carga computacional elevada, inerente tanto à resolução deste sistema como ao cálculo dos respectivos coeficientes de correlação (7-3) e (7-6). Felizmente, é possível optimizar os processamentos referidos através do estudo das características específicas das matrizes utilizadas.

Verificamos facilmente em (7-3) que,

$$\phi_{kl}(i,j) = \phi_{lk}(i,j) \tag{7-8a}$$

e $$\phi_{kl}(i,j) = \phi_{kl}(j,i)\,. \tag{7-8b}$$

Enquanto que a igualdade (7-8a) implica que $\mathbf{R}_{kl} = \mathbf{R}_{lk}$, significando que $\mathbf{R}$ é uma matriz "simétrica por bloco", a segunda igualdade, (7-8b), traduz que cada submatriz

$\mathbf{R}_{lk}$ é por sua vez uma matriz simétrica. Tendo em conta estas duas características, podemos concluir que **R** é também uma matriz simétrica. Por outro lado, como na obtenção dos coeficientes de correlação (7-3) e (7-6) optamos pelo método de covariância, não se verifica qualquer propriedade de simetria adicional nas referidas matrizes [Amir (89)].

Como as funções de base têm suportes temporais finitos, os intervalos de variação dos somatórios no cálculo dos coeficientes de correlação (7-3) e (7-6) diminuem. Considerando a função de base dada em (6-13), concluímos que o intervalo de variação de *n* no somatório (7-3) obtém-se pela intercepção de três intervalos,

$$\begin{cases} 0 \le & n-(k+1-r)m & < mr \\ 0 \le & n-(l+1-r)m & < mr \\ 0 \le & n & \le N-1 \end{cases} \Leftrightarrow \begin{cases} (k+1-r)m & \le n < & (k+1)m \\ (l+1-r)m & \le n < & (l+1)m \\ 0 & \le n \le & N-1 \end{cases} \tag{7-9}$$

Existem duas situações em que a intercepção resulta num intervalo nulo,

$$\begin{aligned} &(k+1-r)m \ge (l+1)m \quad \vee \quad (l+1-r)m \ge (k+1)m \quad \Leftrightarrow \\ &k-l \ge r \quad \vee \quad k-l \le -r \quad \Leftrightarrow \quad |k-l| \ge r \end{aligned} \tag{7-10}$$

Concluímos portanto que nestas condições $\phi_{kl}(i,j)=0$, e por conseguinte as submatrizes $\mathbf{R}_{kl}$ anulam-se sempre que $|k-l| \ge r$. A matriz **R** terá então o seguinte aspecto

$$\mathbf{R} = \underbrace{\begin{bmatrix} \mathbf{R}_{0,0} & \cdots & \mathbf{R}_{0,r-1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{R}_{1,1} & \cdots & \mathbf{R}_{1,r} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{R}_{r-1,0} & \vdots & \mathbf{R}_{2,2} & \cdots & \mathbf{R}_{2,r+1} & & \vdots \\ \mathbf{0} & \mathbf{R}_{r,1} & \vdots & \mathbf{R}_{3,3} & & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_{r+1,2} & & \ddots & & \mathbf{R}_{q-r-1,q} \\ \vdots & \vdots & & & & \mathbf{R}_{q-1,q-1} & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{R}_{q,q-r-1} & \cdots & \mathbf{R}_{q,q} \end{bmatrix}}_{p(q+1)\times p(q+1)} \tag{7-11}$$

Seguindo o mesmo raciocínio, o intervalo de variação de *n* no somatório (7-6) obtém-se pela intercepção de dois intervalos,

$$\begin{cases} 0 \le & n-(k+1-r)m & < mr \\ 0 \le & n & \le N-1 \end{cases} \Leftrightarrow \begin{cases} (k+1-r)m & \le n < & (k+1)m \\ 0 & \le n \le & N-1 \end{cases} \tag{7-12}$$

Porém, há algo extremamente importante que ainda não referimos. Se todo o nosso esforço é verificarmos até que ponto é que um modelo de parâmetros, que

evoluam suavemente, produz resultados positivos, não faz sentido dividirmos o sinal de voz em segmentos e de seguida processarmos cada um deles de uma forma independente dos restantes. Com esse processo, as trajectórias dos parâmetros, que se pretendem suaves, teriam necessariamente descontinuidades entre segmentos. Vejamos então como garantir a continuidade das trajectórias desses parâmetros nas junções dos segmentos.

Em cada dois segmentos adjacentes, devido à sobreposição existente entre *as B-splines* (*B-splines* de ordem $r > 1$), algumas dessas funções encontram-se parcialmente definidas em ambos os segmentos. O mesmo será dizer que as mesmas *B-splines* são responsáveis pela descrição das trajectórias dos parâmetros, quer no fim do primeiro segmento, quer no início do segundo segmento considerado. Do exposto, resta-nos apenas uma saída, visando a garantia da continuidade: quando descrevemos a trajectória de um dado parâmetro ao longo de um segmento a partir das *B-splines*, aquelas que se encontrem definidas apenas parcialmente no início do segmento, devem manter as amplitudes encontradas no segmento anterior. Assim, resulta que os coeficientes $\mathbf{C}_0$, $\mathbf{C}_1$, ..., $\mathbf{C}_{r-2}$ (amplitudes das *B-splines*) associados respectivamente às *B-splines* $b_{-(r-1)}^{(r)}$, $b_{-(r-2)}^{(r)}$, ..., $b_{-1}^{(r)}$ — são estas as que se encontram parcialmente definidas em ambos os segmentos, como se depreende da equação (6-12) — não devem ser calculados, uma vez que os seus valores foram já determinados no segmento anterior. Dessa forma os coeficientes a obter em cada um dos segmentos serão apenas $\mathbf{C}_{r-1}$, $\mathbf{C}_r$, ..., $\mathbf{C}_q$, e por isso, o sistema de matrizes utilizado para o seu cálculo, deixará de ser dado pela equação (7-7), para passar a ser o seguinte

$$\begin{aligned}
&\underbrace{\begin{bmatrix} \mathbf{R}_{r-1,r-1} & \mathbf{R}_{r-1,r} & \cdots & \mathbf{R}_{r-1,q} \\ \mathbf{R}_{r,r-1} & \mathbf{R}_{r,r} & \cdots & \mathbf{R}_{r,q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{q,r-1} & \mathbf{R}_{q,r} & \cdots & \mathbf{R}_{qq} \end{bmatrix}}_{p(q-r)\times p(q-r)} \begin{bmatrix} \mathbf{C}_{r-1}^{*} \\ \mathbf{C}_{r}^{*} \\ \vdots \\ \mathbf{C}_{q}^{*} \end{bmatrix} = \\
&= \begin{bmatrix} \mathbf{P}_{r-1} \\ \mathbf{P}_{r} \\ \vdots \\ \mathbf{P}_{q} \end{bmatrix} - \underbrace{\begin{bmatrix} \mathbf{R}_{r-1,0} & \mathbf{R}_{r-1,1} & \cdots & \mathbf{R}_{r-1,r-2} \\ \mathbf{R}_{r,0} & \mathbf{R}_{r,1} & \cdots & \mathbf{R}_{r,r-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{q,0} & \mathbf{R}_{q,1} & \cdots & \mathbf{R}_{q,r-2} \end{bmatrix}}_{p(q-r)\times p(r-1)} \begin{bmatrix} \mathbf{C}_0 \\ \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_{r-2} \end{bmatrix}
\end{aligned} \tag{7-13}$$

Estando já resolvido o problema da descontinuidade, emerge um outro devido ao aproveitamento, por parte de cada segmento, das amplitudes calculadas para as últimas *B-splines* do segmento anterior. Nesse segmento, as últimas *B-splines* são calculadas entrando em linha de conta apenas com uma parcela inicial dessas funções. Portanto, as *B-splines* assim obtidas ficam optimizadas apenas em relação à parcela contida nesse segmento, não desempenhando o comportamento ideal a restante parcela da *B-spline* que vai ser utilizada no segmento seguinte. De seguida passamos a descrever a forma por nós encontrada de modo a contornarmos esta dificuldade.

Como os valores das amplitudes das últimas *B-splines* (aquelas que se encontram apenas parcialmente definidas) não reflectem o comportamento pretendido para toda a extensão da função, a solução reside em prolongar o segmento de análise, de forma a que as últimas *B-splines* consideradas sejam desprezadas — essas, como se encontram totalmente definidas no segmento seguinte, serão correctamente obtidas nesse segmento. Isto é, devemos introduzir alguma sobreposição nos segmentos de análise. Assim, nas implementações realizadas optamos por introduzir uma sobreposição igual a *rm*, ou seja, no cálculo das amplitudes das *B-splines* consideramos mais *r B-splines*, além das necessárias. Por fim, refira-se que o sistema da equação (7-13) deve ser adaptado de forma a traduzir a alteração proposta. Facilmente se altera de modo a que, para além dos coeficientes válidos, também sejam calculados os coeficientes $\mathbf{C}_{q+1}$, $\mathbf{C}_{q+2}$, ..., $\mathbf{C}_{q+r}$.

De forma a clarificar tudo o que foi dito, segue-se o gráfico ilustrado na Figura 7-1 onde se encontra representada toda a sequência de *B-splines* utilizadas num segmento de análise referente a um exemplo concreto por nós implementado.

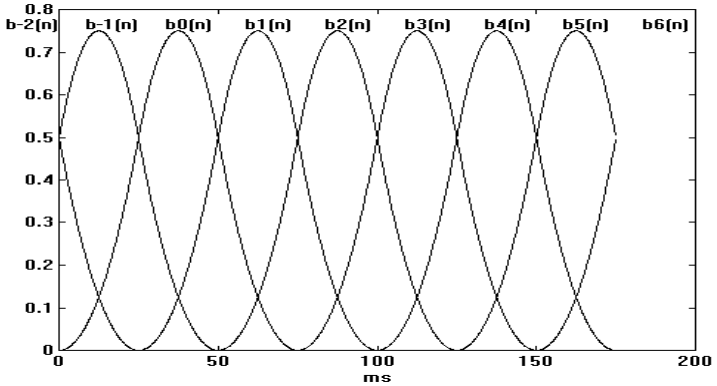


Figura 7-1: *B-splines* de 3ª ordem (*r* = 3).

Neste exemplo modelámos cada um dos parâmetros LP com quatro *B-splines* de 3ª ordem por cada segmento de 100 ms, (r = 3 e q = 5). Note-se que, apenas se pretende obter os conjuntos de coeficientes $\mathbf{C}_2$, $\mathbf{C}_3$, $\mathbf{C}_4$ e $\mathbf{C}_5$, associados respectivamente às *B-splines* $b_0^{(3)}$, $b_1^{(3)}$, $b_2^{(3)}$ e $b_3^{(3)}$. Os valores dos coeficientes $\mathbf{C}_0$ e $\mathbf{C}_1$, associados respectivamente às *B-splines* $b_{-2}^{(3)}$ e $b_{-1}^{(3)}$, foram já obtidos no segmento anterior, e os conjuntos de coeficientes $\mathbf{C}_6$, $\mathbf{C}_7$ e $\mathbf{C}_8$, associados respectivamente às *B-splines* $b_4^{(3)}$, $b_5^{(3)}$ e $b_6^{(3)}$, embora sejam calculados, como as funções se encontram totalmente definidas no segmento seguinte, os seus valores serão desprezados para serem então correctamente obtidos no segmento seguinte.

7.2 Simulação com base no Codificador Standard LPC-10

De forma a testarmos na prática a aplicabilidade da metodologia proposta ao longo deste trabalho, optamos por simular em *Matlab* alguns algoritmos de codificação e descodificação associados a um *vocoder* de predição linear incluindo o modelo LP de parâmetros variáveis. Para o efeito, baseamos a nossa implementação no codificador standard LPC-10 [FS-1015 (84)], com todo o processamento de análise e síntese LP alterado de forma a acomodar o modelo variável.

Nos diagramas das figuras 7-2a e 7-2b encontram-se representados os algoritmos por nós utilizados, respectivamente, na codificação e descodificação.

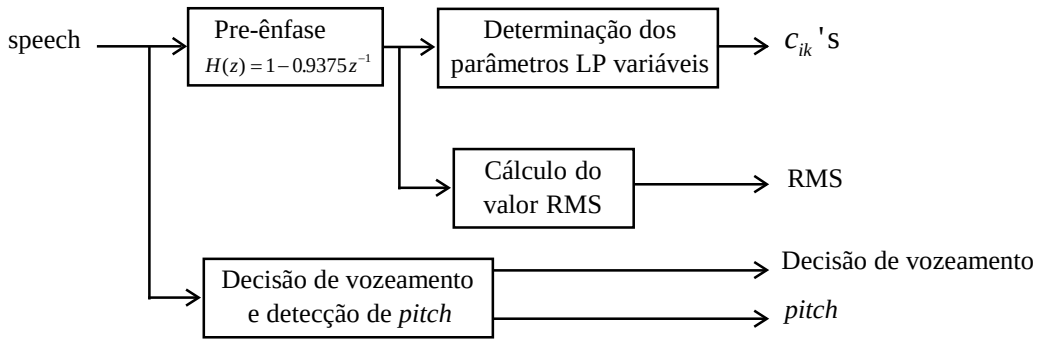


Figura 7-2a: Algoritmo de codificação.

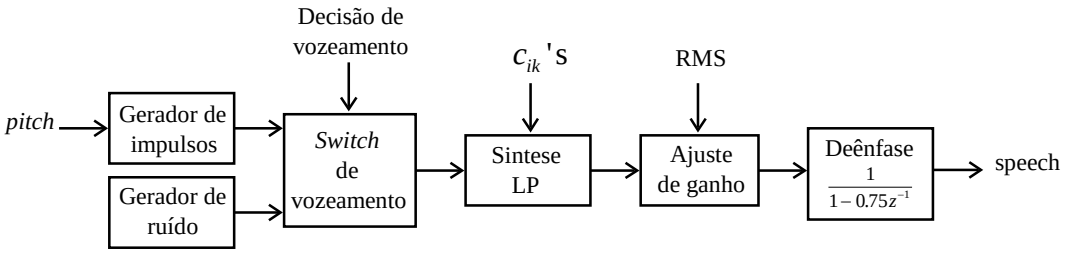


Figura 7-2b: Algoritmo de descodificação.

Para que o algoritmo LPC-10 conduza a um ritmo de transmissão de 2.400 bps — foi para isso que ele foi concebido —, todos os parâmetros extraídos do sinal de voz são eficientemente quantificados — coeficientes LP, ganho RMS, *pitch* e decisão de vozeamento —, de modo serem representados na forma mais compacta possível. Como o nosso objectivo é apenas verificarmos o efeito da modelação LP por parâmetros variáveis, entendemos não sujeitar os parâmetros a qualquer quantificação, de forma a não se introduzir ruído adicional, que poderia mascarar, ou pelo menos desvalorizar, aquele que depende directamente da maneira como se determinam os parâmetros LP.

Em analogia com o LPC-10, todas as implementações por nós realizadas incluem um modelo LP formado por 10 coeficientes. Assim, são sempre 10 os parâmetros a serem modelados pelas *B-splines* (p = 10). Refira-se contudo que, contrariamente ao que acontece com o LPC-10, utilizamos sempre um filtro AR de ordem 10, quer para as zonas vozeadas, quer para as não vozeadas. No LPC-10 é utilizado um filtro AR de quarta ordem para as zonas não vozeadas. Estas duas zonas — vozeadas e não vozeadas — são por nós processadas de mesma forma, exceptuando-se obviamente o modelo de excitação utilizado: as zonas não vozeadas são sintetizadas por um filtro LP excitado por ruído branco; as vozeadas, pelo filtro LP excitado por sequências de impulsos obtidas pela repetição consecutiva do impulso de excitação utilizado no standard LPC-10, representado na Figura 7-3.

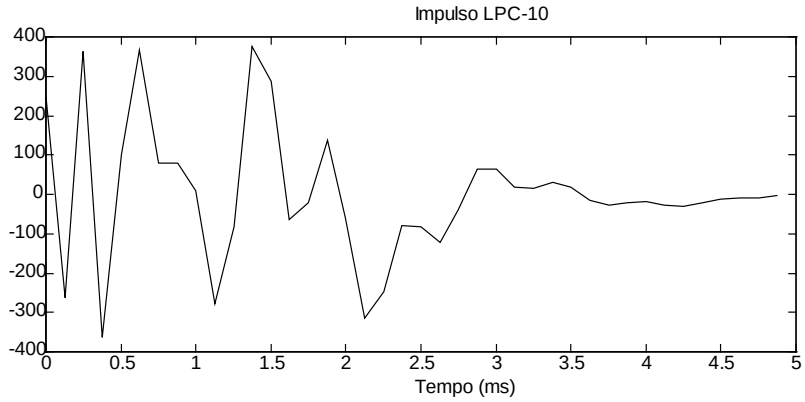


Figura 7-3: Impulso de excitação utilizado no standard LPC-10.

Refira-se por fim que, de forma a explorarmos as potenciais vantagens inerentes à modelação dos parâmetros LP por parte das *B-splines*, teremos necessariamente de utilizar segmentos de análise extremamente longos — deixamos de utilizar segmentos de análise LP de algumas dezenas de milisegundos, para passarmos a utilizar segmentos que poderão chegar a várias centenas de milisegundos. Por isso, como os restantes parâmetros são extraídos da forma convencional, há a necessidade de particionar o sinal de voz em segmentos, e estes por sua vez, em subsegmentos. Passamos então a ter segmentos na ordem das centenas de milisegundos, para extracção dos parâmetros LP, e subsegmentos de apenas 25 ms. para extrair os restantes parâmetros: *pitch*, decisão de vozeamento e amplitude RMS.

7.3 Resultados Obtidos

Comecemos por dizer que, o que seria lógico e normal nesta fase do trabalho, era precisamente servirmo-nos dos critérios de avaliação de desempenho, descritos no capítulo 5, para avaliarmos os vários algoritmos implementados. Porém, por razões que se prendem com dificuldades de ordem prática, ou simplesmente por não fazer sentido a sua utilização nesta aplicação concreta, na maior parte dos testes que realizamos não foi possível seguir nenhum desses critérios.

Quanto aos métodos subjectivos descritos no capítulo 5, embora se tenham realizado testes de audição em todos os sinais codificados, devido à morosidade da codificação, não foi possível quantificar a qualidade perceptual seguindo qualquer um dos métodos propostos. A aplicação implementada envolve processamento

computacional extremamente pesado, chegando a levar várias horas na codificação de algumas centenas de milisegundos de voz. A morosidade de processamento deve-se também ao facto de termos feito a implementação numa linguagem interpretada — embora o *Matlab* esteja optimizado para o processamento vectorial de sinal, cremos que a implementação numa linguagem compilável melhoraria o tempo de processamento.

Por outro lado, não faz sentido utilizarem-se métodos objectivos baseados na relação-sinal-ruído para avaliação de desempenho de *vocoders*, uma vez que a fala obtida por estes codificadores, embora normalmente inteligível, é de qualidade sintética. Exceptuando algumas características, como é o caso da potência e periodicidade, entre outras, a forma de onda de um sinal de voz sintetizado por um *vocoder* nada tem a ver com a do sinal original. A semelhança entre os sinais é visível apenas no domínio da frequência, pois são apenas as suas características espectrais que são reproduzidas por um codificador deste tipo. Por isso, em vez de compararmos o sinais no tempo, resolvemos comparar os sinais no domínio da frequência através da análise de espectrogramas.

O uso de espectrogramas permite-nos visualizar graficamente as características espectrais variáveis no tempo de um sinal de voz. Este método produz uma imagem bidimensional formada por vários níveis de cinzento onde, a dimensão vertical corresponde à frequência e a horizontal ao tempo. Sendo a tonalidade do cinzento proporcional à magnitude do espectro, as frequências ressonantes do tracto vocal (formantes) evidencia-se através de zonas escuras do espectrograma.

Como pretendemos testar apenas as potenciais vantagens inerentes à modelação dos parâmetros LP, optamos por apresentar os resultados referentes apenas a um sinal de voz totalmente vozeado. Mais uma vez, a razão desta opção prende-se com o nosso propósito de evitar que o ruído dependente da maneira como se determinam os parâmetros LP, seja de alguma forma encoberto por ruído inerente a potenciais erros de detecção de *pitch*, ou devido às zonas de transição entre regiões vozeadas e não vozeadas não serem sintetizadas da forma mais correcta pelo modelo utilizado. Por outro lado, como sabemos, a maneira com que se modela o tracto vocal tem uma importância mais determinante precisamente na sintetização dos sons vozeados. O

segmento de fala totalmente vozeado utilizada nos testes encontra-se representado na Figura 7-4a, e o respectivo espectrograma na Figura 7-4b.

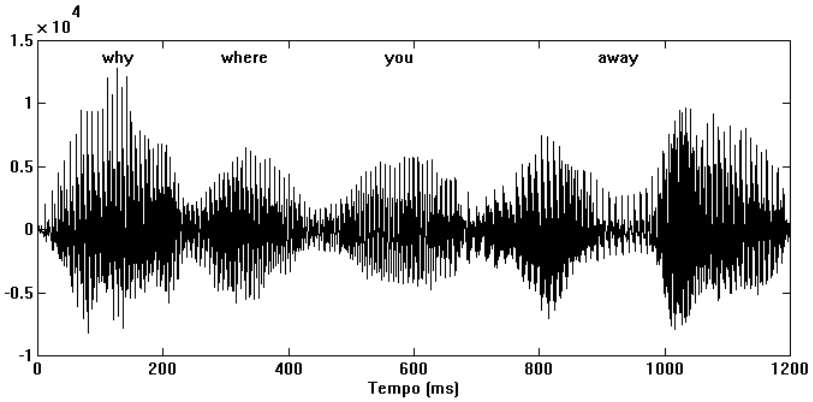


Figura 7-4a: Representação no tempo da frase: "*why where you away*".

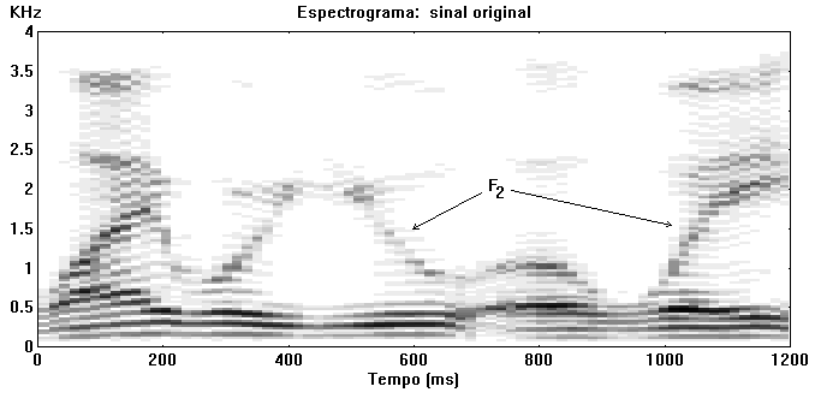


Figura 7-4b: Espectrograma do sinal: "*why where you away*".

Pela análise da Figura 7-4b constata-se que estão representadas no espectrograma as características essenciais do sinal. É bem visível a evolução da segunda formante, assinalada por F_2. Repare-se que a sua trajectória oscila entre cerca de 500 Hz e 2000 Hz. Embora as restantes formantes não se encontrem tão bem definidas, é igualmente possível localizá-las: a F_1 andará entre os 300 e os 700 Hz; a F_3 entre os 1800 e os 2500 Hz; e a F_4 entre os 3200 e os 3600 Hz. Por outro lado, é também visível a existência de vozeamento ao longo do sinal — o vozeamento é caracterizado por riscas horizontais igualmente espaçadas, devido à periodicidade do sinal no tempo.

Uma vez que as *B-splines* de 1ª ordem têm a forma de uma função *rect*() — tal como se depreende da Figura A-10a, em apêndice —, ao modelarmos os parâmetros LP com essas funções, o algoritmo comporta-se como um codificador convencional, isto é, os parâmetros passam a ser actualizados periodicamente, e os seus valores são mantidos fixos ao longo de cada período de tempo correspondente à duração de cada *B-spline* de 1ª ordem. Por conseguinte, sempre que apresentarmos resultados relacionados com o modelo fixo, estes foram obtidos executando o algoritmo por nós desenvolvido baseado em *B-splines* de ordem 1 ($r = 1$).

Uma das formas por nós utilizadas para avaliarmos o desempenho do modelo LP consistiu em quantificar o resíduo de predição — obtido pelo filtro LP inverso excitado com o sinal de voz — através da relação-sinal-ruído. Note-se que este sinal é formado precisamente pelos erros cometidos na predição. Por isso, é de esperar que a SNR do resíduo de predição seja tanto maior, quanto melhor se comportar o modelo LP.

As tabelas a seguir apresentadas contêm os valores das SNR's dos resíduos de predição obtidos para modelos LP baseados em *B-splines* de diferentes ordens e durações.

	SNR (dB)	
Segmento	$r = 1$ ($q = 4$)	$r = 3$ ($q = 6$)
100-150 ms.	4.552	4.513
150-200 ms.	5.819	5.784
200-250 ms.	10.700	10.830
250-300 ms.	14.500	14.400
300-350 ms.	12.010	11.950
350-400 ms.	12.530	12.600
400-450 ms.	16.120	16.000
450-500 ms.	11.240	11.190
500-550 ms.	13.010	13.000
550-600 ms.	8.306	8.346
600-650 ms.	11.730	11.730
650-700 ms.	14.830	14.850
SNR média	11.279 dB	11.266 dB

Tabela 7-1: SNR do resíduo de predição referente a segmentos de 50 ms.

	SNR (dB)			
Segmento	$r = 1$ ($q = 3$)	$r = 2$ ($q = 4$)	$r = 3$ ($q = 5$)	$r = 4$ ($q = 6$)
0-100 ms.	6.404	6.691	6.737	6.709
100-200 ms.	4.679	4.792	4.812	4.809
200-300 ms.	11.020	11.560	11.630	11.520
300-400 ms.	11.670	12.080	12.100	12.090
400-500 ms.	13.240	13.320	13.300	13.300
500-600 ms.	10.290	10.580	10.590	10.580
600-700 ms.	12.110	12.340	12.330	12.330
700-800 ms.	10.320	10.310	10.330	10.310
800-900 ms.	10.160	10.230	10.240	10.230
900-1000 ms.	12.960	13.220	13.300	13.240
1000-1100 ms.	7.131	7.307	7.298	7.317
1100-1200 ms.	8.095	8.166	8.174	8.172
SNR média	9.840 dB	10.050 dB	10.070 dB	10.051 dB

Tabela 7-2: SNR do resíduo de predição referente a segmentos de 100 ms.

Da Tabela 7-1 conclui-se não existir praticamente qualquer diferença entre a SNR obtida com *B-splines* de ordem superior a 1 e a SNR obtida para o modelo fixo ($r = 1$). Da Tabela 7-2 verifica-se já existir alguma melhoria, embora pequena, na SNR referente a *B-splines* de ordem superior a 1 em relação à obtida para o modelo fixo. Esta constatação encontra-se de acordo com o esperado, pois como era suposto, as potencialidades da modelação LP por parâmetros variáveis fazem-se mais sentir quando cada *B-spline* modela um parâmetro por um maior período de tempo. Note-se que, os valores da Tabela 7-1 foram obtidos com uma *B-spline* por cada 10 ms.[1], e os da Tabela 7-2 com uma *B-spline* por cada 25 ms. Além dos resultados apresentados, tentámos igualmente quantificar o resíduo de predição para segmentos superiores a 100 ms. A não inclusão de tabelas com esses valores prende-se com a impossibilidade de codificar todo o sinal, devido ao elevado peso computacional associado ao processamento de segmentos demasiado longos. Conseguiu-se, contudo, obter o resíduo de predição referente a alguns (poucos) segmentos de 200 ms e de 300 ms, ambos modelados por 4 *B-splines*. As SNR's obtidas para estes segmentos confirmaram igualmente a constatação de que a vantagem inerente à utilização de *B-splines* é tanto mais evidente quanto maior for a duração das mesmas.

Vamos, portanto, daqui para a frente realizar todas as simulações com base no modelo LP modelado com quatro *B-splines* por cada segmento de 100 ms.

[1] O número de *B-splines* utilizadas por cada segmento é dado por ($q - r + 2$).

De seguida apresentam-se os espectrogramas referentes a sinais sintetizados a partir de modelos baseados em *B-splines* de diferentes ordens.

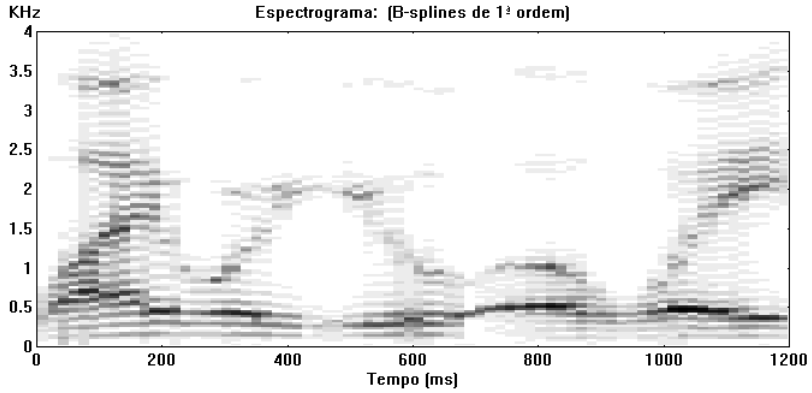


Figura 7-5a: Sinal reconstruído com 4 *B-splines* de 1ª ordem por cada 100 ms.

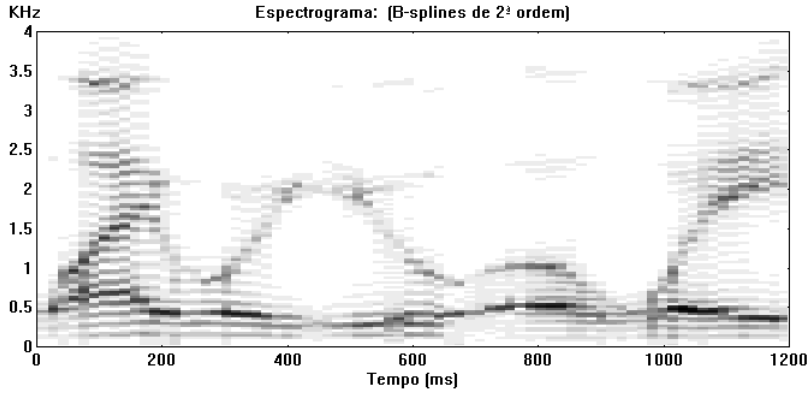


Figura 7-5b: Sinal reconstruído com 4 *B-splines* de 2ª ordem por cada 100 ms.

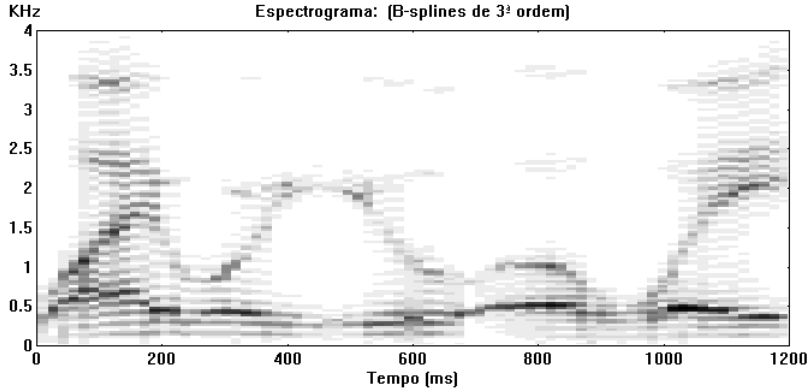


Figura 7-5c: Sinal reconstruído com 4 *B-splines* de 3ª ordem por cada 100 ms.

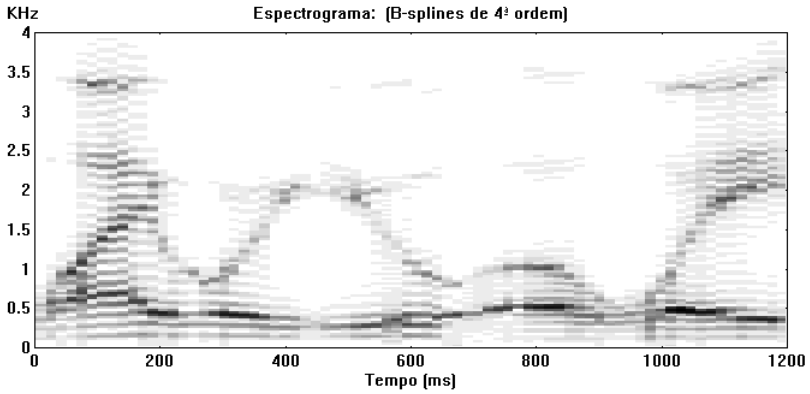


Figura 7-5d: Sinal reconstruído com 4 *B-splines* de 4ª ordem por cada 100 ms.

Pela observação dos espectrogramas, verifica-se que qualquer uma das codificações conseguiu reproduzir as características espectrais de uma forma bastante aceitável, nomeadamente a trajectória da formate F_2. A diferença mais significativa em relação ao sinal original reside na dificuldade em reproduzir a estrutura harmónica relacionada com o *pitch*. No entanto, esta imperfeição nada tem a ver como o modelo LP em estudo. Pode-se ainda verificar a existência de grande semelhança nas trajectórias das formantes F_1, F_3 e F_4 em relação às suas congéneres do sinal original.

Da comparação dos quatro espectrogramas entre si, a conclusão mais importante a retirar é que o modelo LP variável no tempo consegue reproduzir de forma mais correcta a trajectória das formantes. Repare-se que a trajectória F_2 encontra-se melhor definida com os modelos variáveis (figuras 7-5 b, c, d) do que com o modelo fixo (Figura 7-5a). Por outro lado, comparando apenas os espectrogramas dos modelos variáveis entre si, não se verifica qualquer diferença significativa. Contudo, se atendermos apenas à parte da trajectória F_2 após os 1000 ms., encontramos alguma vantagem por parte do modelo com *B-splines* de 3ª ordem em relação aos restantes. Por isso, nas restantes implementações que efectuarmos vamos utilizar modelos com *B-splines* de 3ª ordem.

Refira-se ainda que, em simultâneo com os resultados apresentados, realizaram-se alguns testes de audição. A partir destes verificou-se que os sinais sintetizados são perceptualmente indistinguíveis. No entanto, todos eles evidenciaram fala bastante inteligível, embora de qualidade sintética.

Seguidamente apresentam-se dois conjuntos de gráficos referentes a dois segmentos codificados com 4 *B-splines* de 3ª ordem por cada 100 ms. Os primeiro 6

gráficos (figuras 7-6 a, b, c, d, e, f) referem-se ao segmento de 100 a 200 ms., e os restantes 6 (figuras 7-7 a, b, c, d, e, f) ao segmento de 200 a 300 ms. Cada série de gráficos inclui a representação no tempo do segmento a codificar, a trajectória dos parâmetros LP, o erro cometido na predição, o sinal utilizado na excitação, e por fim, o segmento sintetizado.

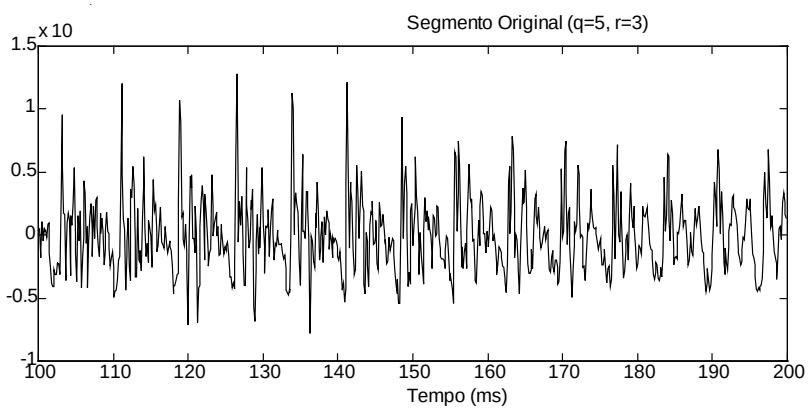


Figura 7-6a: Segundo segmento de sinal a codificar.

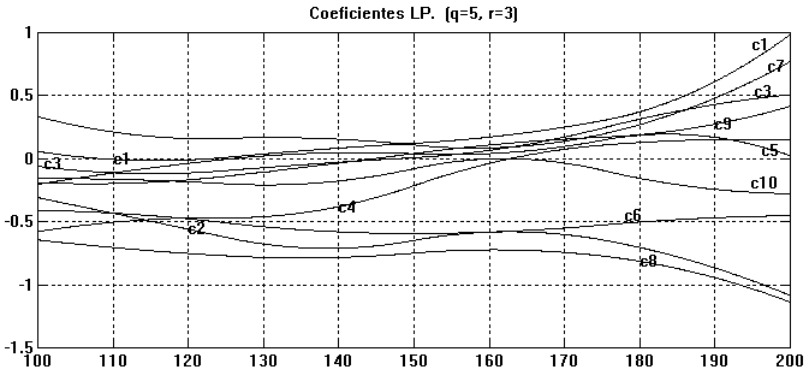


Figura 7-6b: Trajectórias dos coeficientes LP obtidas a partir de *B-splines* de 3ª ordem.

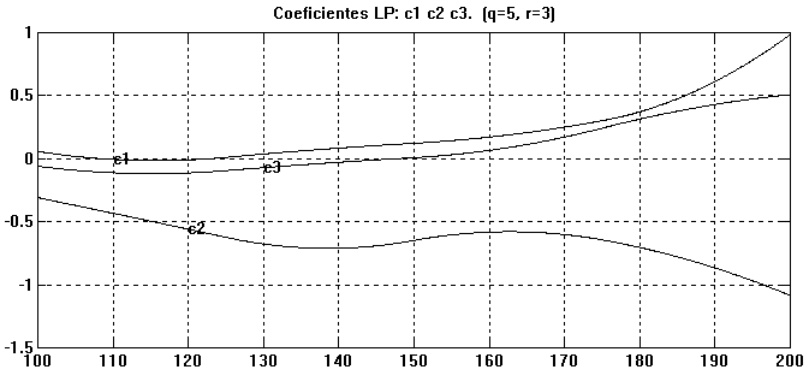


Figura 7-6c: Trajectórias dos 3 primeiros coeficientes LP.

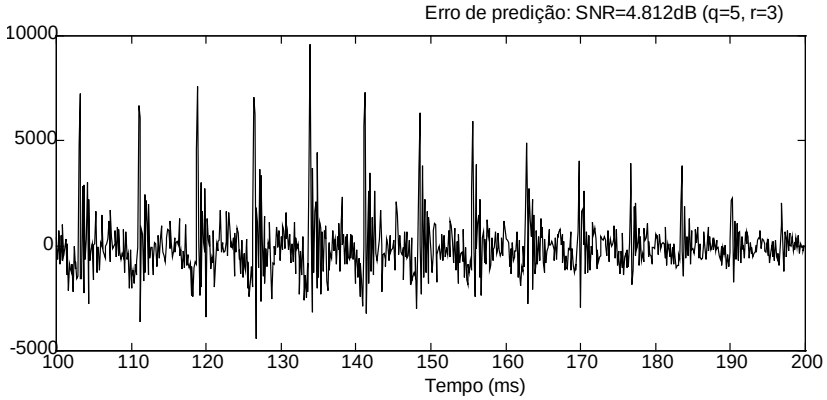


Figura 7-6d: Resíduo de predição.

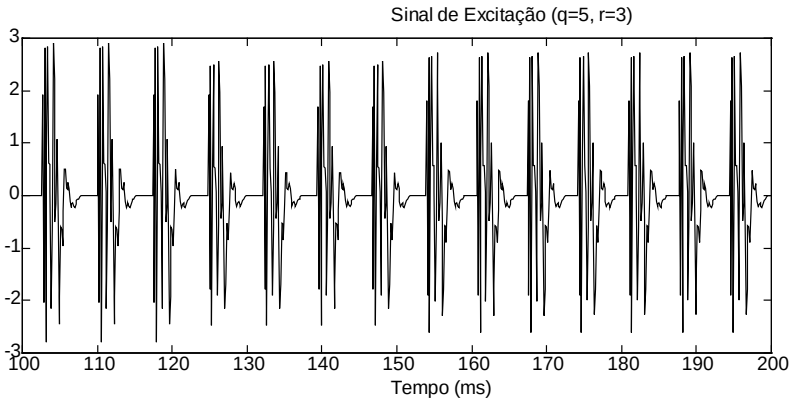


Figura 7-6e: Sinal de excitação usado na síntese.

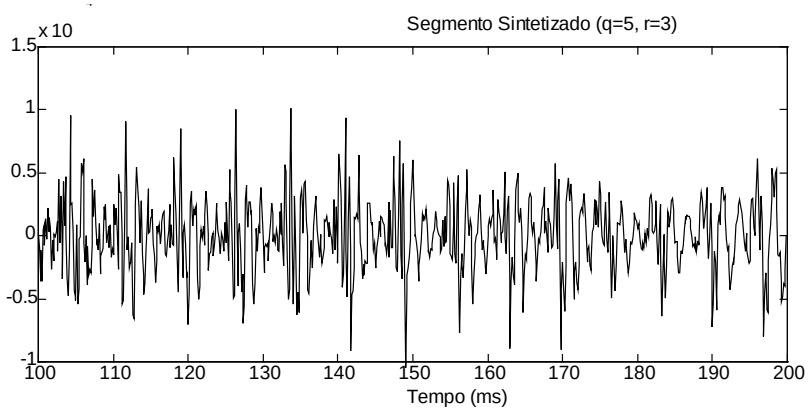


Figura 7-6f: Segmento de voz reconstruído.

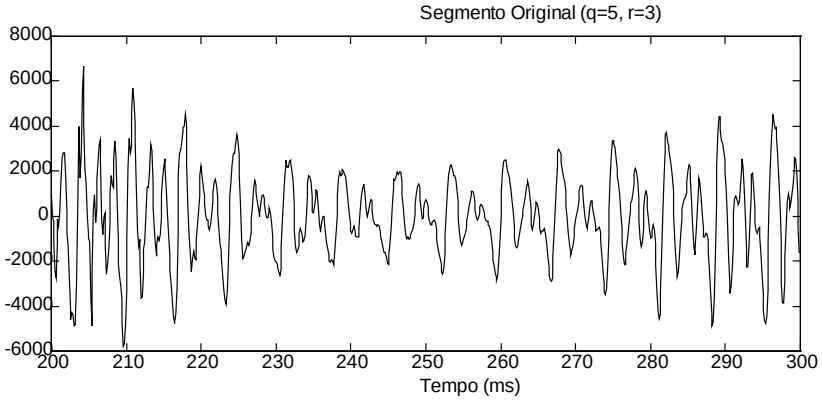


Figura 7-7a: Terceiro segmento de sinal a codificar.

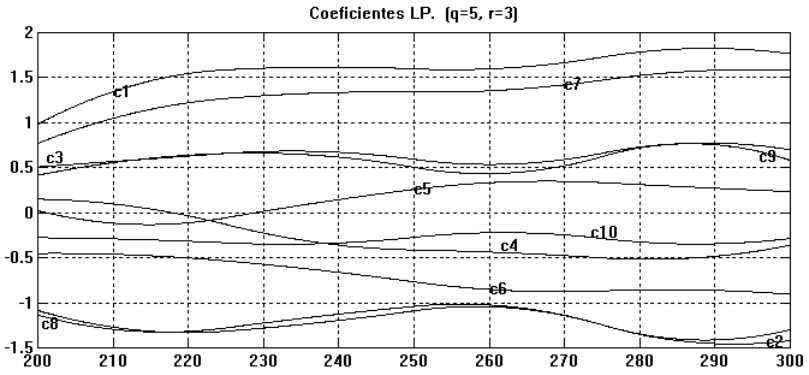


Figura 7-7b: Trajectórias dos coeficientes LP obtidas a partir de *B-splines* de 3ª ordem.

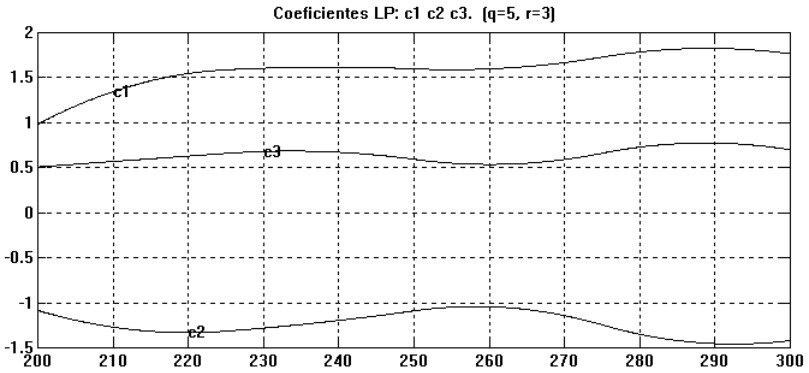


Figura 7-7c: Trajectórias dos 3 primeiros coeficientes LP.

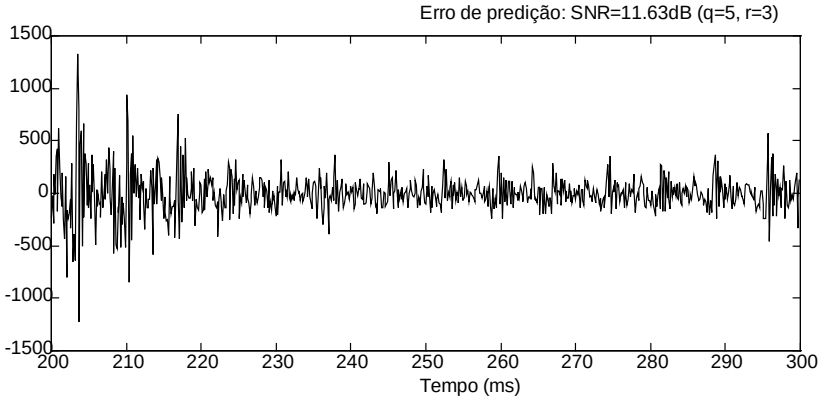


Figura 7-7d: Resíduo de predição.

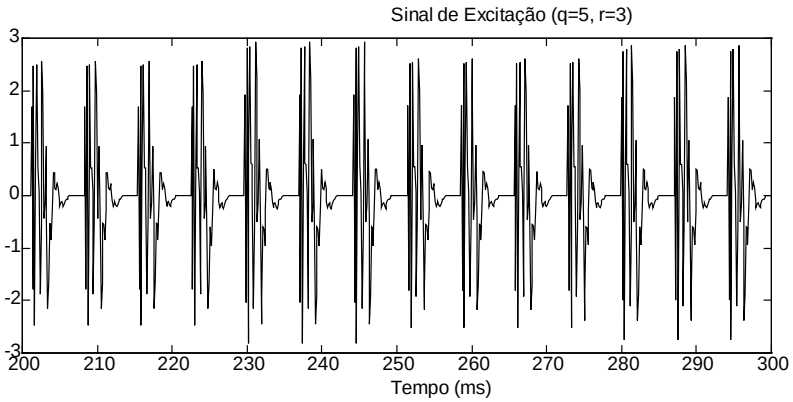


Figura 7-7e: Sinal de excitação usado na síntese.

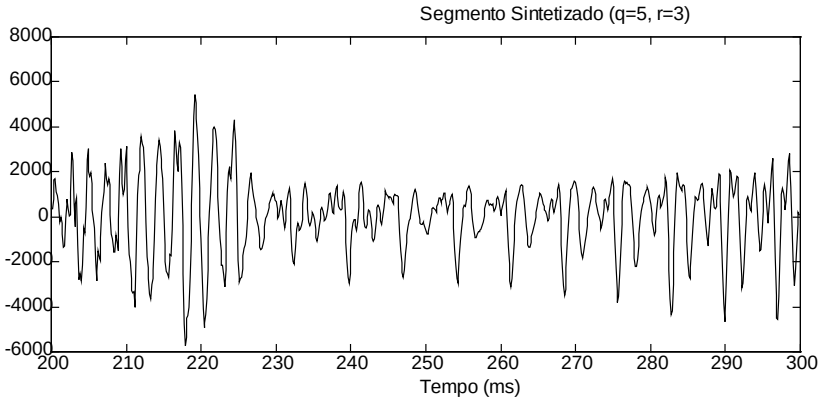


Figura 7-7f: Segmento de voz reconstruído.

Comparando os gráficos das figuras 7-6a e 7-7a respectivamente com os das figuras 7-6f e 7-7f, verifica-se que os sinais sintetizados são semelhantes aos respectivos sinais originais.

Pela análise dos gráficos podemos também comprovar a continuidade da trajectória dos parâmetros LP na junção entre os dois segmentos. Repare-se na evolução dos parâmetros nas figuras 7-6b e 7-7b, e nas figuras 7-6c e 7-7c.

7.4 Comparação com o Modelo LP de Parâmetros Fixos

Com o objectivo de compararmos os resultados obtidos pelo modelo LP variável com os obtidos pelo modelo de parâmetros fixos, apresentamos de seguida o mesmo tipo de gráficos da secção anterior, mas agora referentes à codificação dos dois segmentos de voz pelo modelo convencional de parâmetros fixos.

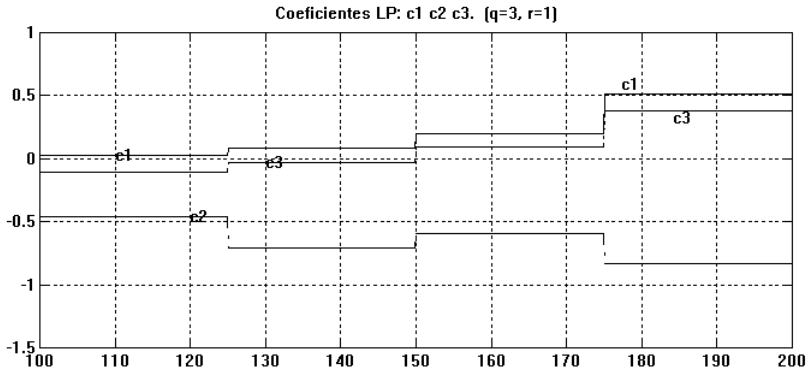


Figura 7-8a: Trajectórias dos 3 primeiros coeficientes do modelo LP convencional, referentes ao segundo segmento de sinal processado.

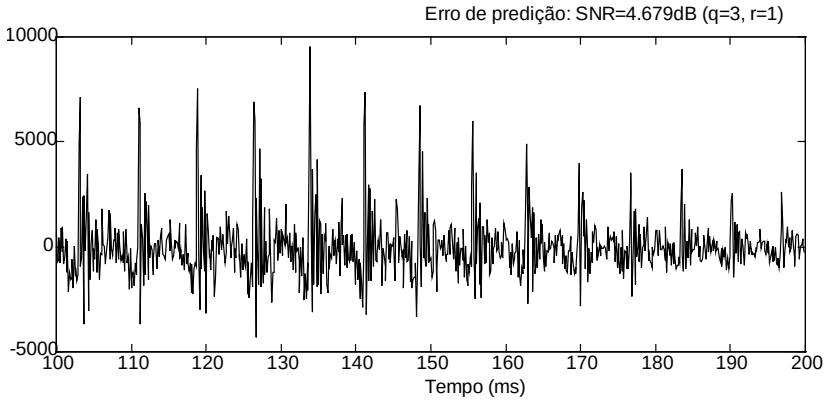


Figura 7-8b: Resíduo de predição.

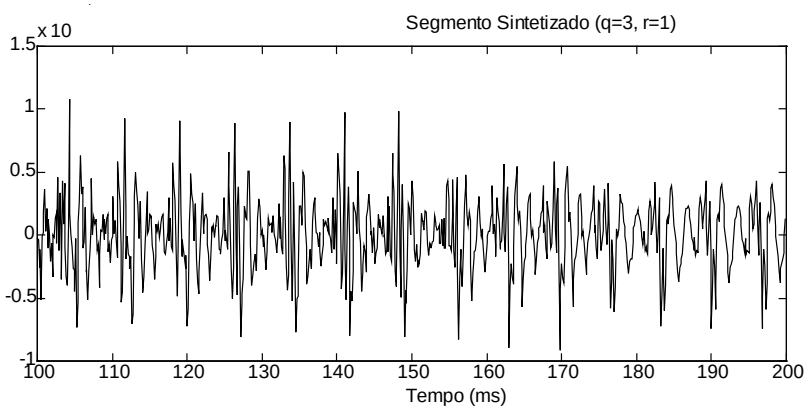


Figura 7-8c: Segmento de voz reconstruído.

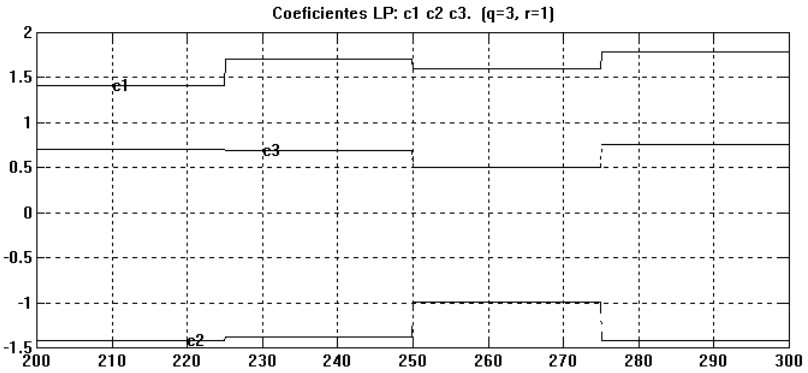


Figura 7-9a: Trajectórias dos 3 primeiros coeficientes do modelo LP convencional, referentes ao terceiro segmento de sinal processado.

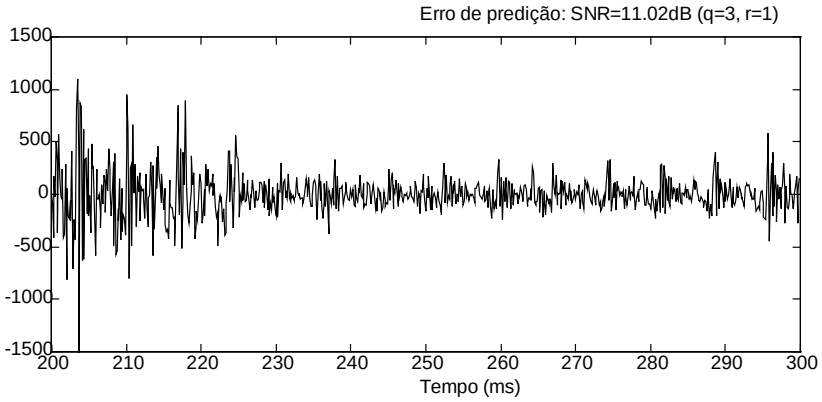


Figura 7-9b: Resíduo de predição.

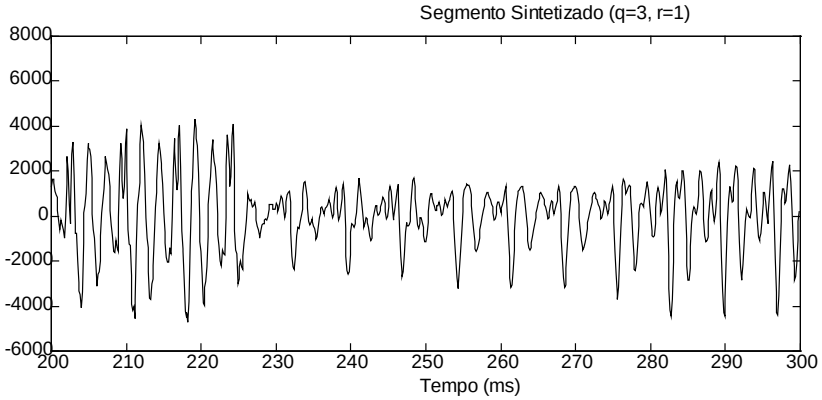


Figura 7-9c: Segmento de voz reconstruído.

Através da comparação entre os sinais sintetizados (figura 7-6f com a 7-8c, e figura 7-7f com a 7-9c) não é possível descortinar qualquer diferença que os distinga em termos de maior ou menor semelhança com os respectivos sinais originais (figuras 7-6a e 7-7a).

Embora também não se verifique grande diferença entre os resíduos de predição (compare-se a figura 7-6d com a 7-8b, e a figura 7-7d com a 7-9b), é possível observar, pelo menos em algumas zonas do segmento, que a potência dos erros cometidos na predição do segmento de 200 a 300 ms. é ligeiramente inferior no modelo variável (figura 7-7d), quando comparado com o modelo fixo (figura 7-9b).

Pela análise dos gráficos das figuras 7-6c e 7-8a, bem como os das figuras 7-7c e 7-9a, podemos comprovar que as trajectórias dos parâmetros variáveis seguem de perto a evolução dos parâmetros fixos. Lembre-se, no entanto, que as referidas

trajectórias dos modelos variáveis não resultam de uma qualquer forma de interpolação dos parâmetros fixos.

Por fim, de forma a ficarmos com uma ideia mais precisa sobre o tipo de evolução associada aos parâmetros LP, resolvemos implementar o modelo LP convencional, actualizando os seus parâmetros ao fim de cada 5 ms. Os gráficos que se seguem ilustram o comportamento desses parâmetros ao longo do segmento de 100 a 200 ms.

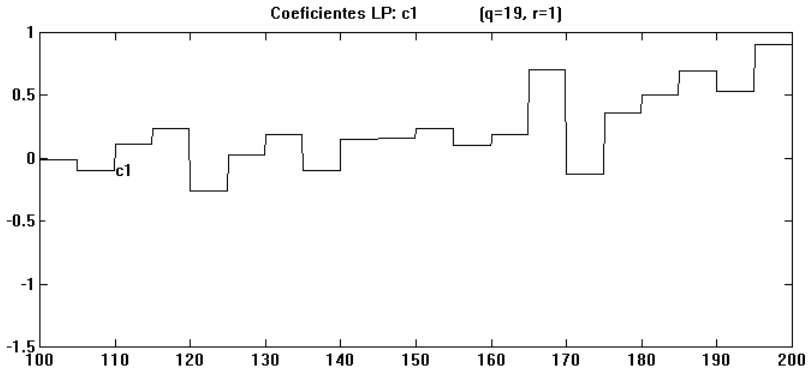


Figura 7-10a: Trajectória do primeiro coeficiente do modelo LP convencional, actualizado em cada 5 ms.

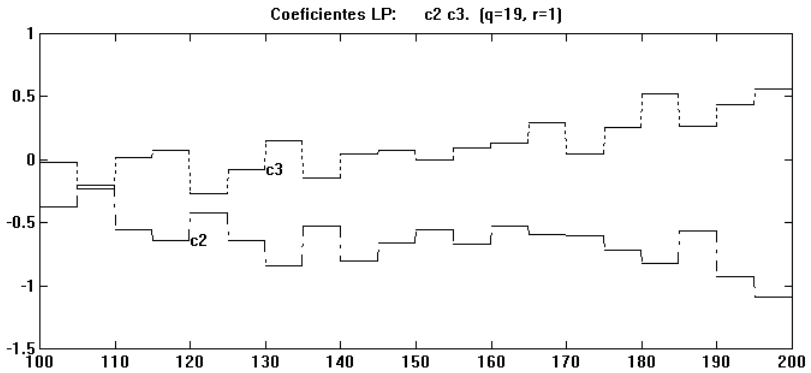


Figura 7-10b: Trajectória do 2º e 3º coeficientes do modelo LP convencional, actualizado em cada 5 ms.

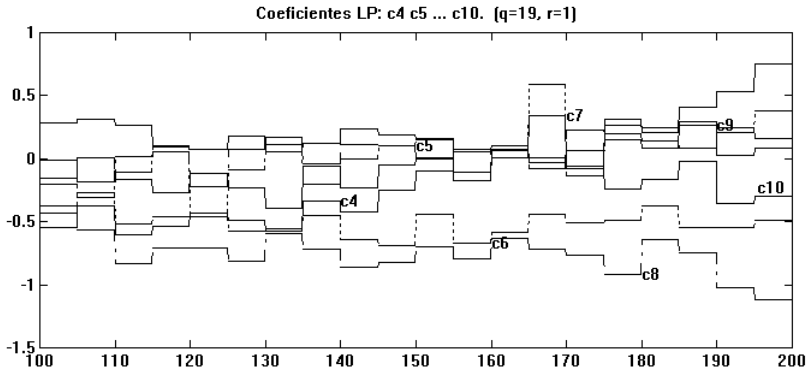


Figura 7-10c: Trajectórias dos restantes coeficientes do modelo LP convencional, actualizado em cada 5 ms.

Como facilmente se verifica, os parâmetros LP não parecem ter uma evolução contínua e suave ao longo do tempo. Talvez isso explique porque razão é que os resultados obtidos pelo modelo variável não tenham sido assim tão diferentes dos obtidos pelo método convencional.

Capítulo 8

Conclusões

8.1 Comentários e Conclusões Finais

Ao propormos neste trabalho um modelo de parâmetros variáveis baseado em *B-splines*, ambicionávamos dessa forma contribuir para o aperfeiçoamento dos algoritmos de codificação de fala. Lamentavelmente, os resultados a que chegamos não nos permitem afirmar que esse objectivo tenha sido alcançado. Na verdade, se alguma diferença existe entre fala sintetizada pelo modelo proposto e a sintetizada pelo modelo convencional, pelo menos em termos de qualidade perceptual — que é a que interessa —, não tem qualquer significado palpável.

O facto de se terem atingido resultados idênticos aos conseguidos por outras técnicas não é encorajador, nomeadamente devido aos vários inconvenientes associados à técnica desenvolvida. Para além do processamento computacional se ter revelado extremamente pesado, existem ainda outros contratempos que convém referir: à codificação está associado um atraso bastante elevado (superior a 100 ms) devido à necessidade de se processarem segmentos extremamente longos; por outro lado, uma vez que os parâmetros LP são variáveis, torna-se impossível anular totalmente a probabilidade de ocorrência de instabilidades no filtro de síntese LP; por fim, na codificação dos parâmetros LP não se podem aplicar as técnicas existentes, extremamente eficientes na quantificação dos coeficientes LP convencionais.

8.2 Trabalho Futuro

A constatação de que a precisão com que se estimam os parâmetros LP é pouco determinante no desempenho de um *vocoder*, não invalida que resultados bem mais

gratificantes possam ser obtidos com outro tipo de codificação, usando a mesma técnica.

Em nosso entender, o mais importante num codificador baseado num modelo LP é a coerência com que se modela a excitação, pois pensamos ser essa a maior fonte de distorção verificada no sinal reproduzido. Uma vez que os codificadores híbridos do tipo análise-por-síntese utilizam modelos de excitação bastante precisos, pensamos que a forma com que se modelam os parâmetros LP pode ter uma maior influência na qualidade da fala sintetizada por esse tipo de codificadores.

Assim, no seguimento deste estudo, achamos que seria interessante tentar aplicar a mesma técnica a um codificador do tipo análise-por-síntese. No entanto, dever-se-á estudar alguma forma de optimizar o processamento, pois foi precisamente devido ao elevado peso computacional associado a esses codificadores que optamos por limitar a simulação do modelo variável a algoritmos do tipo *vocoder*.

Apêndice

Funções de Base *B-spline*

O propósito deste apêndice é fornecer os conceitos teóricos relacionados com as funções *B-spline*, necessários ao desenvolvimento dos modelos variáveis baseados nestas funções. Um tratamento mais exaustivo, e mais elucidativo destes conceitos encontram-se em [Rogers (90)], fonte que serviu de base a este apêndice.

A-1 Introdução

Usando técnicas de interpolação, a forma de onda de uma curva pode ser reconstruída a partir de valores amostrados da curva ideal. Este tipo de técnica pode ser implementada recorrendo a curvas do tipo *spline,* e é caracterizada pelo facto de a curva matemática obtida passar pelos valores amostrados – técnica *"curve fitting"*.

Porém, existem situações em que é de todo interesse que a curva possa ser construída sem qualquer conhecimento à prior da forma da curva desejada. Para implementarmos este tipo de técnica, podemos utilizar curvas *Bézier* ou curvas do tipo *B-spline*. Esta técnica é caracterizada pelo facto de a curva obtida poder passar ou não por alguns dos pontos de controle utilizados na construção da curva – técnica *"curve fairing"*. Mesmo assim, em algumas aplicações, quando desejável, podemos ainda com esta técnica forçar a curva a passar por todos os pontos de controle.

A-2 *Splines* Cúbicas

Embora as *splines* possam descrever curvas tridimensionais, vamos-nos restringir apenas ao espaço bidimensional, pois é aquele que interessa quando se pretende representar a variação da amplitude de um sinal ao longo do tempo.

Uma *spline* física é caracterizada, por exemplo, por uma tira longa e fina de madeira ou plástico usada para descrever uma curva que passe por pontos específicos, como ilustrado na fig. A-1.

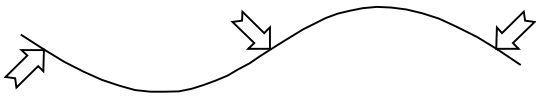


Figura A-1 *Spline* física com 3 pontos de controle.

Sendo a forma da curva resultante determinada pelas posições dos pontos de controle, também conhecidos por "pesos", ao variarmos o número ou as posições dos pesos, a *spline* altera-se, mantendo contudo a suavidade da forma e passando sempre por todos os pontos de controle.

Se se considerar o sistema convencional de coordenadas *XY*, e a variável *y* representar a deflexão da curva ao longo da abcissa *x*, é possível obter a partir da equação de *Euler* [Higdon (67)] a seguinte relação

$$y = B_1 + B_2 x + B_3 x^2 + B_4 x^3. \qquad \text{(A-1)}$$

Onde *y*, representa a deflexão duma curva *spline* do terceiro grau entre dois suportes (ou pontos de controle) consecutivos. Esta expressão mostra tacitamente que a forma da *spline* entre suportes é descrita matematicamente por polinómios do terceiro grau.

Genericamente, uma *spline* é uma curva segmentalmente polinomial de ordem [1] *r* com pontos de junção com derivadas contínuas até à ordem *r*-2. Assim, uma curva do tipo *spline* de ordem *r* pertence a C^{r-2}.

A possibilidade dos segmentos da *spline* corresponderem a polinómios de pequeno grau é bastante útil para fazer interpolação, pois permite reduzir o peso computacional necessário e reduz a ocorrência de instabilidades que normalmente aparecem em curvas de maior grau. Essas instabilidades fazem-se notar através de oscilações indesejáveis que ocorrem, por exemplo, quando vários pontos de controle se encontram dispostos coliniarmente. No entanto, com um polinómio de pequeno grau não é possível fazer passar a curva por um número arbitrário de pontos. São por isso usados segmentos polinomiais adjacentes na construção duma *spline*.

Poder-se-ia escolher o grau dos segmentos polinomiais de forma a que cada segmento abrangesse três ou mais pontos. No entanto, como o grau do polinómio a

utilizar aumenta com o número de pontos a abranger por cada segmento, torna-se preferível, por razões já mencionadas, que cada segmento polinomial abranja apenas dois pontos. Para este caso a *spline* cúbica parece ser a melhor escolha, pois é a curva de menor grau que permite um ponto de inflexão e conduz a curvaturas suaves.

No caso de considerarmos que a curva representa um sinal, se em (A-1) fizermos corresponder à ordenada *y* uma amplitude *P*(*t*) e à abcissa *x* o parâmetro *t*, obtemos a seguinte equação análoga referente a um segmento duma *spline* cúbica

$$P(t) = B_1 + B_2 t + B_3 t^2 + B_4 t^3, \quad para\ t_1 \le t \le t_2, \tag{A-2}$$

onde t_1 e t_2 são os valores da abcissa referentes aos instantes inicial e final do segmento, e *P*(*t*) é a amplitude num qualquer ponto da curva entre esses instantes. Existem, portanto, quatro coeficientes constantes B_i a serem especificados necessariamente a partir de outras tantas "condições de fronteira".

Impondo um determinado declive em ambas as extremidades do segmento, conjuntamente com os valores da curva *P*(*t*) que deve ter nesses mesmos pontos, obtemos as condições necessárias para a obtenção dos coeficientes em (A-2). Como ilustrado na fig. A-2, P_1 e P_2 representam a amplitude (*ordenada*) da curva em cada uma das extremidades do segmento, e P_1' e P_2' os declives dos respectivos vectores tangentes, nesses mesmos pontos.

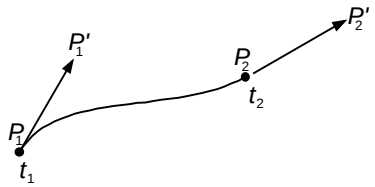


Figura A-2 Segmento de uma *Spline* cúbica.

Diferenciando a função *P*(*t*) em ordem a *t*, obtém-se

$$P'(t) = B_2 + 2B_3 t + 3B_4 t^2, \quad para\ t_1 \le t \le t_2, \tag{A-3}$$

e assim, atribuindo às funções $P(t)$ e $P'(t)$ os valores das extremidades, ficamos na posse das quatro condições necessárias para obtenção das constantes em (A-2),

$$P(t_1) = P_1; \tag{A-4a}$$

$$P(t_2) = P_2; \tag{A-4b}$$

[1] grau $r-1$.

$$P'(t_1) = P_1'; \tag{A-4c}$$

$$P'(t_2) = P_2'. \tag{A-4d}$$

Isto é, ficamos com quatro equações para quatro incógnitas B_i's.

De modo a simplificarmos podemos assumir, sem perda de generalidade, que $t_1 = 0$. Logo

$$P(0) = B_1 = P_1; \tag{A-5a}$$

$$P'(0) = B_2 = P_1'; \tag{A-5b}$$

$$P(t_2) = B_1 + B_2 t_2 + B_3 t_2^2 + B_4 t_2^3 = P_2; \tag{A-5c}$$

$$P'(t_2) = B_2 + 2B_3 t_2 + 3B_4 {t_2}^2 = P_2'. \tag{A-5d}$$

Resolvendo em ordem a B_3 e B_4, obtém-se

$$B_3 = \frac{3(P_2 - P_1)}{t_2^2} - \frac{2P_1'}{t_2} - \frac{P_2'}{t_2}; \tag{A-6a}$$

$$B_4 = \frac{2(P_1 - P_2)}{t_2^3} + \frac{P_1'}{t_2^2} + \frac{P_2'}{t_2^2}. \tag{A-6b}$$

Como os valores de B_1, B_2, B_3 e B_4 obtidos determinam totalmente um segmento duma *spline* cúbica, logicamente podemos afirmar que a forma do segmento de curva depende apenas das localizações dos dois extremos e dos respectivos vectores tangentes nesses pontos. Substituindo na equação (A-2), os coeficientes encontrados em (A-5) e (A-6), obtemos a equação geral para um segmento duma *spline* cúbica,

$$P(t) = P_1 + P_1' \cdot t + \left[\frac{3(P_2 - P_1)}{t_2^2} - \frac{2P_1'}{t_2} - \frac{P_2'}{t_2}\right] t^2 + \left[\frac{2(P_1 - P_2)}{t_2^3} + \frac{P_1'}{t_2^2} + \frac{P_2'}{t_2^2}\right] t^3. \tag{A-7}$$

Portanto, neste momento, estamos na posse da fórmula para construir a curva de interpolação entre dois pontos consecutivos. Mantém-se, contudo, a necessidade de estipular os valores P_1' e P_2', que contêm o declive dos vectores tangentes nos pontos extremos do segmento.

Para se construir uma curva completa, faz-se a junção de múltiplos segmentos, de forma a que, o segundo ponto de controle do primeiro segmento, de cada dois consecutivos, coincida com o primeiro ponto de controle do segundo segmento. A fig.A-3 ilustra este tipo de junção

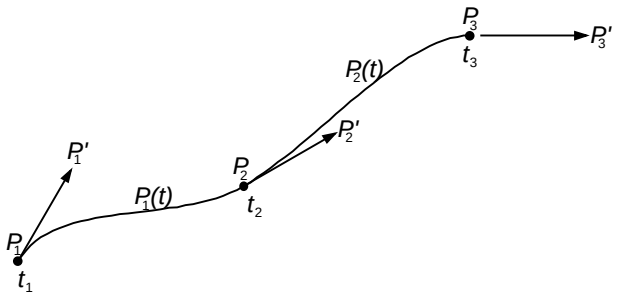


Figura A-3 Dois segmentos adjacentes de uma *Spline* cúbica.

Tendo em conta que o ponto (t_2, P_2) e respectivo declive P_2' são comuns aos dois segmentos, pode-se desde já aplicar a expressão (A-7) a cada um dos segmentos de modo a se obterem os correspondentes segmentos de curva. Embora os declives extremos P_1' e P_3' sejam impostos à partida, o declive do ponto intermédio P_2' não é conhecido. Existe, contudo, a possibilidade de o determinar, impondo uma condição de continuidade nesse ponto.

Se considerarmos $P(t)$ a função que representa a concatenação de todos os segmentos que formam a curva *spline,* isto é, $P(t) = P_k(t)$ com $t_k \le t \le t_{k+1}$, pelo exposto facilmente se verifica que tanto $P(t)$ como $P'(t)$ na fig.A-3 são funções contínuas no ponto de junção — fig.A-3. Se impusermos igualmente a continuidade de $P''(t)$, de modo a podermos determinar P_2', estamos simultaneamente a garantir que $P(t)$ pertença a C^2, tal como já referido anteriormente. Ou seja, $P(t)$ é uma *spline* cúbica ($r = 4$), e como tal pertence a C^{r-2}. Assim, depois de diferenciarmos a equação (A-3), vem para cada um dos segmentos

$$P_1''(t) = 2B_3^{(1)} + 6B_4^{(1)}t, \quad para \;\; 0 \le t \le t_2 \tag{A-8a}$$

$$P_2''(t) = 2B_3^{(2)} + 6B_4^{(2)}(t - t_2), \quad para \;\; t_2 \le t \le t_3. \tag{A-8b}$$

Igualando estas duas funções no ponto de junção obtemos a condição que faltava,

$$P_1''(t_2) = P_2''(t_2) \quad \Leftrightarrow \quad 2B_3^{(1)} + 6B_4^{(1)}t_2 = 2B_3^{(2)}.$$

Substituindo agora os coeficientes pelos respectivos valores das equações (A-6), resulta

$$\left(\frac{3(P_2-P_1)}{t_2^2}-\frac{2P_1'}{t_2}-\frac{P_2'}{t_2}\right)+3\left(\frac{2(P_1-P_2)}{t_2^3}+\frac{P_1'}{t_2^2}+\frac{P_2'}{t_2^2}\right)t_2 =$$
$$=\left(\frac{3(P_3-P_2)}{(t_3-t_2)^2}-\frac{2P_2'}{(t_3-t_2)}-\frac{P_3'}{(t_3-t_2)}\right)$$

Multiplicando de seguida ambos os termos por $(t_3-t_2)t_2$ e resolvendo em ordem a P_2', obtém-se

$$P_2'=\frac{t_2}{2t_3}\left(\frac{3(P_3-P_2)}{(t_3-t_2)}-P_3'\right)+\frac{t_3-t_2}{2t_3}\left(\frac{3(P_2-P_1)}{t_2}-P_1'\right), \qquad \text{(A-9)}$$

que é precisamente o declive do vector tangente que desconhecíamos.

Neste momento encontra-se totalmente definida a curva resultante da concatenação de dois segmentos duma *spline* cúbica, ilustrada na fig.A-3. Agora, facilmente generalizamos estes resultados. De modo a encontrarmos as expressões que descrevam a curva constituída por *n*-2 pontos de controle intermédios, 2 pontos extremos e *n*-1 segmentos, necessitamos de impor os declives nos extremos da curva, garantir a passagem da curva pelos pontos de controle, e garantir a existência da continuidade C^2 em todos os pontos intermédios. Usando a notação utilizada na fig.A-4, as especificações da curva obedecem aos seguintes requisitos:

- $P(t_k)=P_k, \quad para \;\; 1\le k\le n;$ (A-10a)
- $P_{k-1}'(t_k)=P_k'(t_k), \quad para \;\; 2\le k\le n-1;$ (A-10b)

  $P_1'(t_1)=P_1';$ (A-10c)

  $P_{n-1}'(t_n)=P_n';$ (A-10d)
- $P_{k-1}''(t_k)=P_k''(t_k), \quad para \;\; 2\le k\le n-1;$ (A-10e)

onde $P(t)$ representa a curva total constituída por todos os segmentos, e $P_{k-1}(t)$ e $P_k(t)$ correspondem a quaisquer dois segmentos adjacentes, tal como ilustrado na fig.A-4.

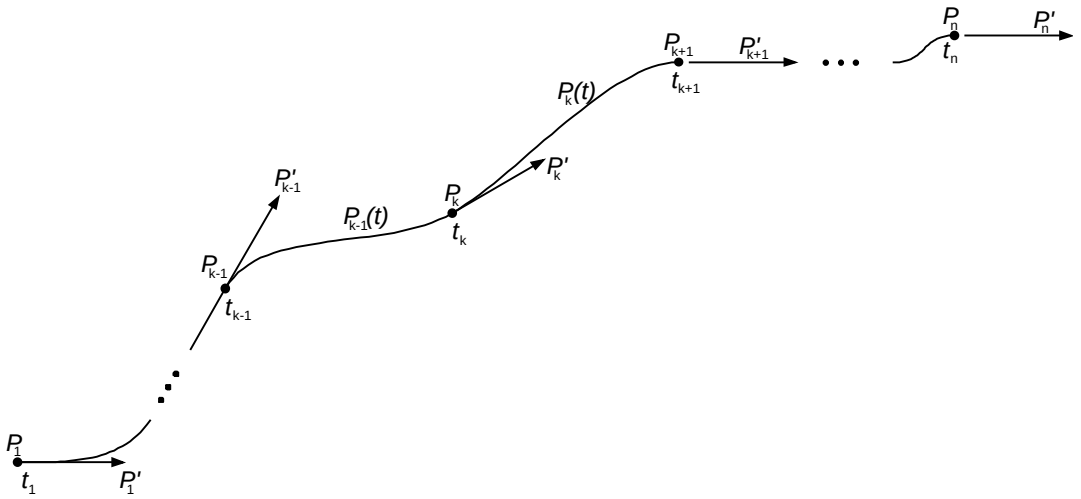


Figura A-4 *Spline* cúbica com *n* pontos de controle.

A partir da expressão (A-7), podemos deduzir a equação análoga que descreve a forma de um segmento genérico $P_k(t)$,

$$P_k(t) = P_k + P_k' \cdot (t - t_k) + \left[\frac{3(P_{k+1} - P_k)}{(t_{k+1} - t_k)^2} - \frac{2P_k'}{(t_{k+1} - t_k)} - \frac{P_{k+1}'}{(t_{k+1} - t_k)} \right](t - t_k)^2 + \\ + \left[\frac{2(P_k - P_{k+1})}{(t_{k+1} - t_k)^3} + \frac{P_k'}{(t_{k+1} - t_k)^2} + \frac{P_{k+1}'}{(t_{k+1} - t_k)^2} \right](t - t_k)^3 \tag{A-11}$$

Esta expressão foi obtida impondo as condições (A-10a), (A-10b), (A-10c) e (A-10d). A *spline* é constituída por $n-1$ segmentos ($k = 1,2,\ldots,n-1$), encontrando-se o k ésimo segmento definido em $t_k \le t \le t_{k+1}$. Resta, portanto, impor a continuidade da segunda derivada nos pontos de junção entre cada par de segmentos adjacentes (condição A-10e), isto é

$$P_{k-1}''(t_k) = P_k''(t_k) \quad \Leftrightarrow \quad 2B_3^{(k-1)} + 6B_4^{(k-1)}(t_k - t_{k-1}) = 2B_3^{(k)}. \tag{A-12}$$

Adaptando agora as expressões (A-6) de modo a traduzirem as especificações referentes aos segmentos intermédios pretendidos, podemos substituir os coeficientes constantes da equação (A-12) pelos seus valores. Feitas as substituições e depois algumas simplificações, vem

$$(t_{k+1} - t_k)P_{k-1}' + 2(t_{k+1} - t_{k-1})P_k' + (t_k - t_{k-1})P_{k+1}' = \\ = 3\left[\frac{t_k - t_{k-1}}{t_{k+1} - t_k}(P_{k+1} - P_k) + \frac{t_{k+1} - t_k}{t_k - t_{k-1}}(P_k - P_{k-1}) \right] \tag{A-13}$$

Ficamos, portanto, na presença de uma equação de três incógnitas, respectivamente P'_{k-1}, P'_k e P'_{k+1}, não sendo por isso possível resolvê-la directamente. Mas ao garantirmos a continuidade da segunda derivada em todos os pontos intermédios da curva $(k = 2,3,\ldots,n-1)$, obtemos um sistema representado pela seguinte equação matricial

$$\begin{bmatrix} t_3 - t_2 & 2(t_3 - t_1) & (t_2 - t_1) & 0 & \cdots & 0 \\ 0 & t_4 - t_3 & 2(t_4 - t_2) & (t_3 - t_2) & 0 & \vdots \\ \vdots & & & \ddots & & 0 \\ 0 & \cdots & 0 & t_n - t_{n-1} & 2(t_n - t_{n-2}) & (t_{n-1} - t_{n-2}) \end{bmatrix} \times$$

$$\times \begin{bmatrix} P'_1 \\ P'_2 \\ \vdots \\ P'_n \end{bmatrix} = \begin{bmatrix} 3\frac{t_2 - t_1}{t_3 - t_2}(P_3 - P_2) + 3\frac{t_3 - t_2}{t_2 - t_1}(P_2 - P_1) \\ 3\frac{t_3 - t_2}{t_4 - t_3}(P_4 - P_3) + 3\frac{t_4 - t_3}{t_3 - t_2}(P_3 - P_2) \\ \vdots \\ 3\frac{t_{n-1} - t_{n-2}}{t_n - t_{n-1}}(P_n - P_{n-1}) + 3\frac{t_n - t_{n-1}}{t_{n-1} - t_{n-2}}(P_{n-1} - P_{n-2}) \end{bmatrix} \qquad \text{(A-14a)}$$

ou, $\tilde{\mathbf{M}} \times \mathbf{P}' = \tilde{\mathbf{R}}$. (A-14b)

Uma vez que existem apenas $n-2$ equações para n incógnitas (declive de n *vectores tangente,* $\mathbf{P}'$), a matriz $\tilde{\mathbf{M}}$ não é quadrada, e por conseguinte não pode ser invertida de modo a podermos encontrar os valores de $\mathbf{P}'$, ou seja, o sistema é indeterminado. É precisamente para ultrapassar esta indeterminação que se impõem à partida os declives nos extremos da curva – *declives* P'_1 e P'_n. Desta forma, o sistema passa a ser representado por

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & & 0 \\ t_3 - t_2 & 2(t_3 - t_1) & (t_2 - t_1) & 0 & & \vdots \\ 0 & t_4 - t_3 & 2(t_4 - t_2) & (t_3 - t_2) & 0 & \\ \vdots & & & \ddots & & 0 \\ & & 0 & t_n - t_{n-1} & 2(t_n - t_{n-2}) & (t_{n-1} - t_{n-2}) \\ 0 & \cdots & & 0 & 0 & 1 \end{bmatrix}$$

$$\times \begin{bmatrix} P_1' \\ P_2' \\ \vdots \\ P_n' \end{bmatrix} = \begin{bmatrix} P_1' \\ 3\frac{t_2 - t_1}{t_3 - t_2}(P_3 - P_2) + 3\frac{t_3 - t_2}{t_2 - t_1}(P_2 - P_1) \\ \vdots \\ 3\frac{t_3 - t_2}{t_4 - t_3}(P_4 - P_3) + 3\frac{t_4 - t_3}{t_3 - t_2}(P_3 - P_2) \\ 3\frac{t_{n-1} - t_{n-2}}{t_n - t_{n-1}}(P_n - P_{n-1}) + 3\frac{t_n - t_{n-1}}{t_{n-1} - t_{n-2}}(P_{n-1} - P_{n-2}) \\ P_n' \end{bmatrix} \quad \text{(A-15a)}$$

ou, $\mathbf{M} \times \mathbf{P}' = \mathbf{R}$. (A-15b)

Agora, a matriz **M** é quadrada, podendo por isso ser invertida, e sendo uma matriz de diagonal dominante[2] é não singular o que significa que a sua inversão conduz a uma solução única. A matriz **M** é também tridiagonal[3], o que permite reduzir o peso computacional requerido na sua inversão.

Os declives em todos os pontos da curva *spline* serão então dados por

$$\mathbf{P}' = \mathbf{M}^{-1} \times \mathbf{R}. \quad \text{(A-16)}$$

Depois de conhecidos todos os declives, os coeficientes dos segmentos *spline* são facilmente obtidos através da generalização das equações (A-5a), (A-5b) e (A-6). Deste modo, os coeficientes do *k* ésimo segmento (*com* $k = 1,2,\ldots,n-1$) são dados por

$$B_1^{(k)} = P_k; \quad \text{(A-17a)}$$

$$B_2^{(k)} = P_k'; \quad \text{(A-17b)}$$

$$B_3^{(k)} = \frac{3(P_{k+1} - P_k)}{(t_{k+1} - t_k)^2} - \frac{2P_k'}{(t_{k+1} - t_k)} - \frac{P_{k+1}'}{(t_{k+1} - t_k)}; \quad \text{(A-17c)}$$

[2] Numa matriz de diagonal dominante, a amplitude de cada coeficiente da diagonal principal excede a amplitude de todos os coeficientes sobre a mesma linha.

[3] Uma matriz tridiagonal tem apenas coeficientes na diagonal principal, primeira diagonal inferior e primeira diagonal superior.

$$B_4^{(k)} = \frac{2\left(P_k - P_{k+1}\right)}{\left(t_{k+1} - t_k\right)^3} + \frac{P'_k}{\left(t_{k+1} - t_k\right)^2} + \frac{P'_{k+1}}{\left(t_{k+1} - t_k\right)^2}. \qquad \text{(A-17d)}$$

Recapitulando, para gerar uma curva *spline* através de *n* pontos ($P_1, P_2, \cdots, P_n$), com vectores tangentes de declives P'_1 e P'_n nos extremos, começa-se por recorrer à equação (A-16) para determinação de todos os vectores tangentes ($P'_2, P'_3, \cdots, P'_{n-1}$) nos pontos intermédios de controle. Seguidamente para cada segmento *k*, determinam-se os coeficientes $B_i^{(k)}$'s ($1 \le i \le 4$) a partir das amplitudes (P_k e P_{k+1}) e declives (P'_k e P'_{k+1}) nos respectivos extremos do segmento, servindo-nos para o efeito, das equações (A-17). Por fim, adaptando a equação (A-2), obtém-se a expressão que traduz totalmente a forma da curva referente a um qualquer segmento *k*,

$$P_k(t) = B_1^{(k)} + B_2^{(k)}\left(t - t_k\right) + B_3^{(k)}\left(t - t_k\right)^2 + B_4^{(k)}\left(t - t_k\right)^3, \quad t_k \le t \le t_{k+1} \qquad \text{(A-18)}$$

Esta equação coincide com a já apresentada em (A-11), na qual os coeficientes $B_i^{(k)}$'s encontram-se substituídos pelos seus valores. Pegando nessa equação e fazendo a mudança de variável $\tau = \dfrac{t - t_k}{t_{k+1} - t_k}$, obtém-se a função $P_k(\tau)$ que, depois de algum processamento de modo a pôr em evidência P_k, P_{k+1}, P'_k e P'_{k+1}, resulta na seguinte expressão

$$\begin{aligned} P_k(\tau) = &\left[1 - 3\tau^2 + 2\tau^3\right]P_k + \left[3\tau^2 - 2\tau^3\right]P_{k+1} + \\ &+\left[\tau - 2\tau^2 + \tau^3\right]\left(t_{k+1} - t_k\right)P'_k + \left[-\tau^2 + \tau^3\right]\left(t_{k+1} - t_k\right)P'_{k+1} \end{aligned} \qquad \text{(A-19)}$$

$P_k(\tau)$ descreve o *k* ésimo segmento da curva *spline*, em função do parâmetro τ que traduz percentualmente a posição relativa do instante *t* no segmento. Nesta equação os parâmetros P_k, P_{k+1}, P'_k e P'_{k+1} contêm a informação geométrica da curva e os factores a eles associados são conhecidos por funções *Blending* (ou *Weighting Functions*). Existem, portanto, quatro funções *Blending*. Respectivamente

$$F_1^{(k)}(\tau) = 1 - 3\tau^2 + 2\tau^3; \qquad \text{(A-20a)}$$

$$F_2^{(k)}(\tau) = 3\tau^2 - 2\tau^3; \qquad \text{(A-20b)}$$

$$F_3^{(k)}(\tau) = \left[\tau - 2\tau^2 + \tau^3\right]\left(t_{k+1} - t_k\right); \qquad \text{(A-20c)}$$

$$F_4^{(k)}(\tau) = \left[-\tau^2 + \tau^3\right]\left(t_{k+1} - t_k\right). \qquad \text{(A-20d)}$$

Estas funções cúbicas contêm a informação referente ao peso da contribuição de cada um dos parâmetros P_k, P_{k+1}, P'_k e P'_{k+1}, na formação da *k* ésima curva. Ou seja, qualquer ponto da curva $P_k(\tau)$ é uma soma pesada dos parâmetros P_k, P_{k+1}, P'_k e P'_{k+1}, associados aos extremos do segmento, com os pesos dados pelas funções *Blending*,

$$P_k(\tau) = F_1^{(k)}(\tau)\cdot P_k + F_2^{(k)}(\tau)\cdot P_{k+1} + F_3^{(k)}(\tau)\cdot P'_k + F_4^{(k)}(\tau)\cdot P'_{k+1} \qquad \text{(A-21)}$$

Como em qualquer segmento *k* se verifica $t_k \le t \le t_{k+1}$, então entre quaisquer par de pontos de controle consecutivos $0 \le \tau \le 1$. Na fig.A-5 encontram-se representadas as funções *Blending* de uma *spline* cúbica com espaçamento entre pontos de controle unitário ($t_{k+1} - t_k = 1$).

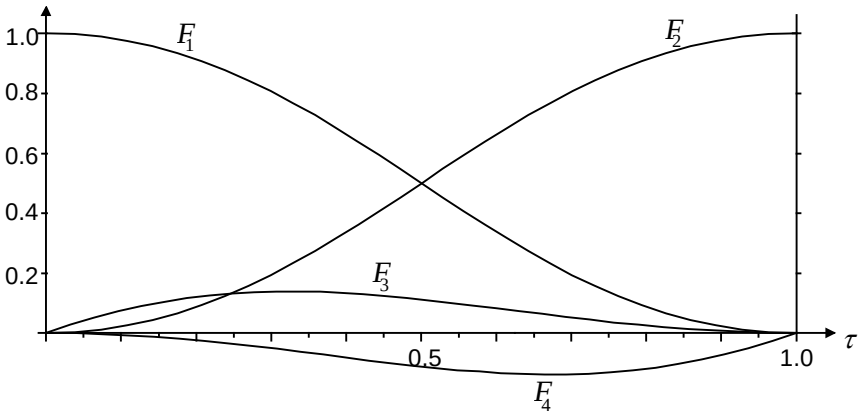


Figura A-5 Funções *Blending* de uma *Spline* cúbica.

Como se pode observar, no inicio do segmento, $F_1^{(k)}(0) = 1$ e $F_2^{(k)}(0) = F_3^{(k)}(0) = F_4^{(k)}(0) = 0$. Deste modo, facilmente se verifica através da equação (A-21) que a curva passa pelo ponto de controle P_k. Da mesma forma se demonstra que a curva passa por P_{k+1}, pois $F_2^{(k)}(1) = 1$ e $F_1^{(k)}(1) = F_3^{(k)}(1) = F_4^{(k)}(1) = 0$.

Note-se a relação de amplitudes que existe entre as funções $F_1^{(k)}$ e $F_2^{(k)}$ e as funções $F_3^{(k)}$ e $F_4^{(k)}$ no caso de considerarmos o espaçamento entre amostras unitário. Esta diferença significativa de amplitudes mostra que os valores P_k e P_{k+1} têm maior influência na formação da curva do que os declives P'_k e P'_{k+1}. É igualmente importante referir a simetria que existe entre $F_1(\tau)$ e $F_2(\tau)$, e entre $F_3^{(k)}(\tau)$ e $F_4^{(k)}(\tau)$, onde se verifica, respectivamente, $F_2(\tau) = F_1(1-\tau)$ e

$F_4^{(k)}(\tau) = -F_3^{(k)}(1-\tau)$. Além disso, a soma de $F_1(\tau)$ com $F_2(\tau)$ é sempre unitária, logo $F_2(\tau) = 1 - F_1(\tau)$. Saliente-se, por fim, que apenas as duas funções *Blending* $F_3^{(k)}$ e $F_4^{(k)}$ dependem do espaçamento amostral $(t_{k+1} - t_k)$, por conseguinte a influência dos declives P_k' e P_{k+1}' é tanto maior quanto maior for a distância $(t_{k+1} - t_k)$. Esta característica permite que as funções se adaptem à distância amostral, quando esta for variável.

Até aqui assumimos que os declives P_1' e P_n' dos vectores tangentes nos extremos da curva *spline* eram conhecidos. No entanto, se nem todos os pontos de controle são dados ou existir a necessidade de controlar a forma da curva nos extremos, pode-se recorrer a condições de fronteira alternativas. Uma das alternativas é, por exemplo, especificar a curvatura em ambos os extremos da curva. Podemos impor curvaturas nulas nesses pontos. Ou seja,

$$P''(t_1) = P_1''(0) = 0 \quad \Leftrightarrow \quad 2B_3^{(1)} = 0 \tag{A-22a}$$

$$P''(t_n) = P_{n-1}''(t_n) = 0 \quad \Leftrightarrow \quad 2B_3^{(n-1)} + 6B_4^{(n-1)}\left(t_n - t_{n-1}\right) = 0 \tag{A-22b}$$

Substituindo o coeficiente dado pela expressão (A-6a) na equação (A-22a), obtém-se

$$P_1' + \frac{1}{2}P_2' = \frac{3\left(P_2 - P_1\right)}{2t_2} \tag{A-23}$$

A primeira linha do sistema de matrizes (A-15) é, portanto, alterada para

$$\begin{bmatrix} 1 & 1/2 & 0 & \cdots & 0 \end{bmatrix} \times \begin{bmatrix} P_1' \end{bmatrix} = \left\lceil \frac{3\left(P_2 - P_1\right)}{2t_2} \right\rceil .$$

De forma análoga, adaptando a expressão (A-6b) de modo a traduzir as especificações referentes ao último segmento da curva, podemos substituir os coeficientes constantes da equação (A-22b) pelos seus valores. Feitas as substituições e após algumas simplificações, vem

$$P_{n-1}' + 2P_n' = \frac{3\left(P_n - P_{n-1}\right)}{t_n - t_{n-1}}, \tag{A-24}$$

de onde se obtém a última linha do sistema de matrizes

$$\begin{bmatrix} 0 & \cdots & 0 & 1/2 & 1 \end{bmatrix} \times \begin{bmatrix} P_n' \end{bmatrix} = \left\lfloor \frac{3\left(P_n - P_{n-1}\right)}{2\left(t_n - t_{n-1}\right)} \right\rfloor .$$

Duas outras condições de fronteira alternativas são as condições *cíclica* e *anticíclica*. A condição *cíclica* serve, por exemplo, para produzir curvas periódicas, e é especificada impondo que os dois extremos da curva tenham o mesmo declive e a mesma curvatura, isto é,

$$P_1'(0) = P_n'(t_n)\,, \tag{A-25a}$$

$$P_1''(0) = P_n''(t_n)\,. \tag{A-25b}$$

Na curva *anticíclica* as condições a impor são

$$P_1'(0) = -P_n'(t_n)\,, \tag{A-26a}$$

$$P_1''(0) = -P_n''(t_n)\,. \tag{A-26b}$$

Aplicando um processo semelhante ao utilizado nas equações (A-22), encontrávamos facilmente as alterações a implementar na primeira e última linhas do sistema de matrizes $\mathbf{M} \times \mathbf{P}' = \mathbf{R}$.

A-3 Curvas *Bézier*

Na técnica anteriormente apresentada a curva passa através dos pontos de controle ou amostras. Sendo uma técnica do tipo *"curve fitting"*, a sua utilização destina-se essencialmente para descrever formas e contornos a partir valores amostrados, obtidos experimentalmente ou através de cálculos matemáticos. Contudo, pode ser desejável ou mesmo necessário construir a curva sem conhecimento à prior de quaisquer pontos por onde a curva deva passar — técnica *"curve fairing"*. *Pierre Bézier* desenvolveu um método alternativo de representação de curvas que se enquadra nesta segunda técnica.

Uma curva *Bézier* é determinada a partir de um polígono formado por um conjunto de pontos de controle, tal como ilustrado na fig.A-6.

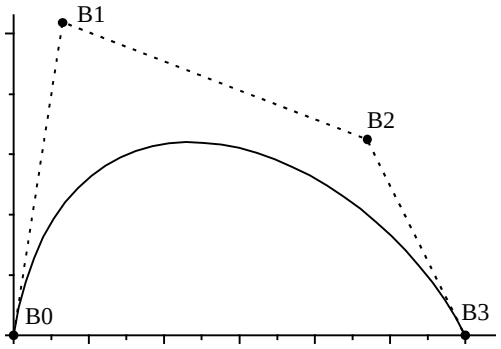


Figura A-6 Curva *Bézier* e respectivo polígono de construção.

Pode-se desde já enumerar um conjunto de propriedades associadas a este tipo de curvas:

- as funções de base são reais;
- o grau do segmento de curva polinomial é inferior em uma unidade ao número de pontos que definem o polígono de construção;
- a curva polinomial tende a seguir a forma do polígono de construção;
- o primeiro e último pontos do polígono coincidem respectivamente com o inicio e o fim do segmento de curva;
- os vectores tangentes nos extremos da curva têm a direcção do primeiro e último lados do polígono, respectivamente;
- a curva encontra-se totalmente inserida no interior do maior polígono convexo que é possível obter a partir dos vértices do polígono de construção;
- a curva nunca oscila em torno de qualquer linha recta mais vezes que o respectivo polígono de construção.

De modo a evidenciar a relação existente entre os polígonos e as curvas deles resultantes, encontram-se representados na fig.A-7 dois polígonos de quatros vértices e as respectivas curvas *Bézier* que, dado o número de vértices, correspondem a polinómios do terceiro grau.

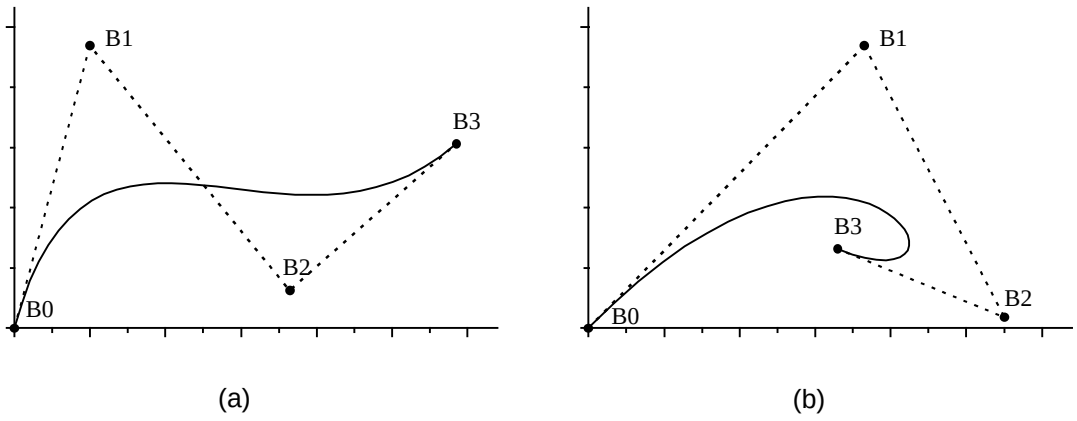


Figura A-7 Curvas *Bézier* cúbicas.

Para polígonos de $n+1$ vértices, a formula geral da curva *Bézier* de grau *n* é dada por

$$P(\tau) = \sum_{k=0}^{n} B_k J_k^{(n)}(\tau) \quad 0 \le \tau \le 1, \tag{A-27}$$

onde a *k* ésima base de *Bézier* de grau *n* ou função *Blending* é dada por

$$J_k^{(n)}(\tau) = \binom{n}{k} \tau^k (1-\tau)^{n-k}, \tag{A-28}$$

com $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$.

Nesta expressão assume-se que $(0)^0 \equiv 0$ e $0! \equiv 1$.

Portanto, $J_k^{(n)}$ é a *k* ésima função *Blending* a utilizar na formação de uma curva definida por um polígono de $n+1$ vértices, existindo no total tantas funções quantos os vértices. Na fig.A-8 encontram-se representadas as funções *Blending* de segundo e terceiro grau, a serem utilizadas em polígonos, respectivamente, de três e quatro vértices. O valor máximo da *k* ésima função encontra-se em $\tau = k/n$, sendo *n*, com já mencionado, o grau da função.

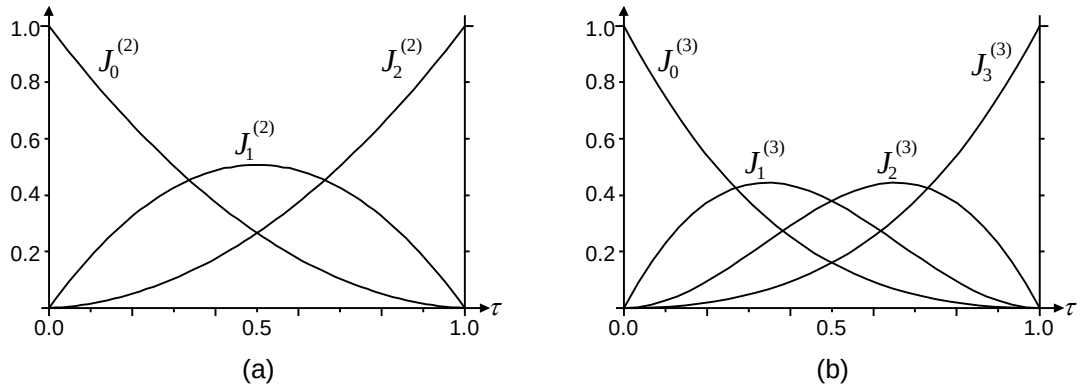


Figura A-8 Funções *Bézier*: (a) Funções *Blending* do segundo grau; (b) Funções *Blending* do terceiro grau.

Resolvendo a equação (A-27) para $\tau = 0$ e $\tau = 1$, facilmente se verifica que $P(0) = B_0$ e $P(1) = B_n$, constatando-se desta forma que os extremos da curva coincidem com os do polígono. A fig.A-8 ilustra igualmente este resultado. Pode-se demonstrar, também, que o somatório de todas as funções *Blending* é sempre igual a um, isto é,

$$\sum_{k=0}^{n} J_k^{(n)}(\tau) = 1 \quad 0 \le \tau \le 1. \tag{A-29}$$

Embora não seja necessário especificar os declives nos extremos de uma curva *Bézier* isolada, quando se pretende concatenar várias curvas deste tipo torna-se necessário ter conhecimento sobre a primeira e segunda derivadas nos extremos da curva para que se possa manter a continuidade do declive e da curvatura nos pontos de junção.

É possível demostrar que

$$P'(0) = n(B_1 - B_0), \tag{A-30a}$$

$$P'(1) = n(B_n - B_{n-1}), \tag{A-30b}$$

$$P''(0) = n(n-1)(B_0 - 2B_1 + B_2), \tag{A-31a}$$

$$P''(1) = n(n-1)(B_n - 2B_{n-1} + B_{n-2}). \tag{A-31b}$$

Pela análise das equações (A-30) verifica-se que os vectores tangentes no primeiro e último pontos da curva têm os declives respectivamente do primeiro e último lados do polígono. Da mesma forma, constata-se através das equações (A-31) que a segunda derivada nos mesmos pontos depende apenas dos três vértices mais próximos de cada extremo.

Se ao ligarmos duas curvas *Bézier* de grau n , as identificarmos por $P_1(t)$ e $P_2(t)$ com vértices, respectivamente $B_k^{(1)}$ e $B_k^{(2)}$, a continuidade do declive no ponto de junção é dada por

$$P_1'(1) = P_2'(0). \tag{A-32}$$

Substituindo os dois termos da equação pelos valores das equações (A-30), vem

$B_n^{(1)} - B_{n-1}^{(1)} = B_1^{(2)} - B_0^{(2)}$.

Para haver continuidade da curva no ponto de junção, o último e primeiro vértices respectivamente do primeiro e segundo polígonos têm que ser coincidentes, logo $B_0^{(2)} = B_n^{(1)}$ e portanto

$$B_n^{(1)} - B_{n-1}^{(1)} = B_1^{(2)} - B_n^{(1)}. \tag{A-33}$$

Deste resultado conclui-se que a continuidade do declive implica que os três vértices $B_{n-1}^{(1)}$, $B_n^{(1)} \equiv B_0^{(2)}$ e $B_1^{(2)}$ sejam colineares. Como as duas curvas são da mesma ordem, o vértice intermédio, que coincide com o ponto de junção, encontra-se a meia distância entre os outros dois. A fig.A-9 ilustra a disposição dos vértices dos dois polígonos referentes a duas curvas *Bézier* do terceiro grau adjacentes.

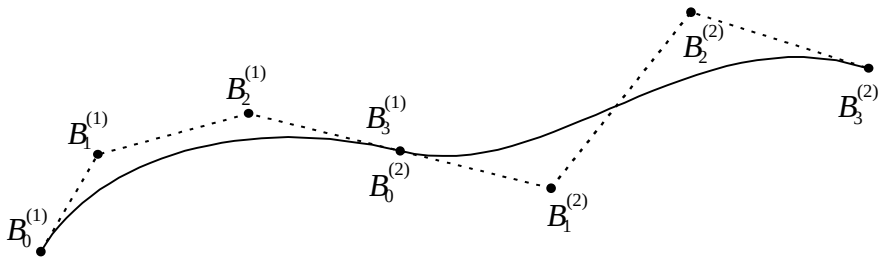


Figura A-9 Continuidade da primeira derivada entre duas curvas *Bézier* adjacentes.

Até este momento apenas garantimos a continuidade C^1 no ponto de junção. De modo a garantirmos a continuidade C^2 torna-se necessário impor a continuidade da segunda derivada em toda a extensão da curva.

A continuidade da segunda derivada no ponto de junção é dada por

$$P_1''(1) = P_2''(0). \tag{A-34}$$

Substituindo os dois termos da equação pelos valores das equações (A-31), vem

$$B_n^{(1)} - 2B_{n-1}^{(1)} + B_{n-2}^{(1)} = B_0^{(2)} - 2B_1^{(2)} + B_2^{(2)}.$$

Utilizando novamente as condições da continuidade em C^1, obtém-se

$$B_2^{(2)} = B_{n-2}^{(1)} + 4(B_n^{(1)} - B_{n-1}^{(1)}). \quad \text{(A-35)}$$

Como se verifica, consegue-se exprimir o terceiro vértice do segundo polígono em função dos três últimos vértices do primeiro polígono. Deste modo, os cinco vértices $B_{n-2}^{(1)}$, $B_{n-1}^{(1)}$, $B_n^{(1)} \equiv B_0^{(2)}$, $B_1^{(2)}$ e $B_2^{(2)}$ formam um polígono convexo.

Todos estes requisitos introduzem demasiadas restrições no projecto de uma curva deste tipo. Na prática, de modo a garantir a continuidade da segunda derivada, é necessário utilizar, quase sempre, curvas polinomiais de um grau bastante elevado.

A-4 *B-splines*

A criação de uma curva a partir dos vértices de um polígono de construção pressupõe a utilização de funções de base que de alguma forma estabeleçam uma aproximação entre o polígono e a curva pretendida. No entanto, como já mencionado, a utilização das funções de base do tipo *Bézier* limitam a flexibilidade da curva resultante. Neste método, o número de vértices utilizados estabelece a ordem polinomial da curva, por conseguinte, a única maneira de alterar a ordem da curva é alterar o número de vértices do polígono. Por exemplo, polígonos de quatro vértices implicam necessariamente curvas de grau três, e vice-versa. Uma outra limitação é o facto de as funções de base *Bézier* terem um comportamento global, isto é, as funções $J_k^{(n)}(\tau)$ na equação (A-27) nunca se anulam em toda a extensão da curva. Uma vez que qualquer ponto da curva resulta de uma média ponderada de todos os vértices (equação A-27), qualquer alteração num desses vértices altera a curva em toda a sua extensão. Portanto esta característica inviabiliza por completo a possibilidade de introduzir qualquer alteração no interior da curva com efeito localizado.

Uma base *B-spline*[4] é uma outra base de funções, que tem a base de *Bézier* como caso particular, não tendo no entanto as suas limitações. Esta base tem geralmente um comportamento não global, que significa que cada vértice exerce, por intermédio da função *B-spline* a si associada, uma influência local na curva reproduzida, não afectando por isso a forma da curva em pontos distantes da zona de influência. A zona de influência de um vértice é precisamente a extensão da curva onde a função *B-spline* a ele associada é diferente de zero.

As funções *B-spline* permitem, igualmente, obter uma curva de qualquer grau inferior ao número de vértices utilizados. Portanto, permite alterar o grau da curva resultante sem modificar o número de vértices do polígono. A teoria de funções *B-spline* foi inicialmente sugerida por *Schoenberg* e posteriormente foi utilizada por *Gordon* e *Riesenfeld* na definição de curva.

A partir de agora sempre que falarmos em curva, deverá ser interpretada como representando a forma de um sinal no tempo. Deste modo, os vértices passam a corresponder a pontos de controle ou a amostras e as coordenadas deixam de traduzir distâncias geométricas. A abcissa passa a traduzir a evolução no tempo e a ordenada a amplitude do sinal.

A $k^{\acute{e}\ sima}$ função *B-spline* normalizada de ordem *r*, $F_k^{(r)}$, encontra-se definida pela formula recursiva de Cox-deBoor [Rogers (90)] do seguinte modo

$$F_k^{(1)}(t) = \begin{cases} 1 & \text{se } t_k \le t < t_{k+1} \\ 0 & \text{caso contrario} \end{cases}, \text{ e} \qquad \text{(A-36a)}$$

$$F_k^{(r)}(t) = \frac{t - t_k}{t_{k+r-1} - t_k} F_k^{(r-1)}(t) + \frac{t_{k+r} - t}{t_{k+r} - t_{k+1}} F_{k+1}^{(r-1)}(t), \quad para\ r > 1, \qquad \text{(A-36b)}$$

onde t_k é o instante correspondente ao inicio da função de base $F_k^{(r)}(t)$. Os valores $t_0, t_1, \cdots$ poderão ou não encontrar-se igualmente espaçados. Na equação (A-36b) convenciona-se que $0/0 = 0$.

[4] *Basis Spline*

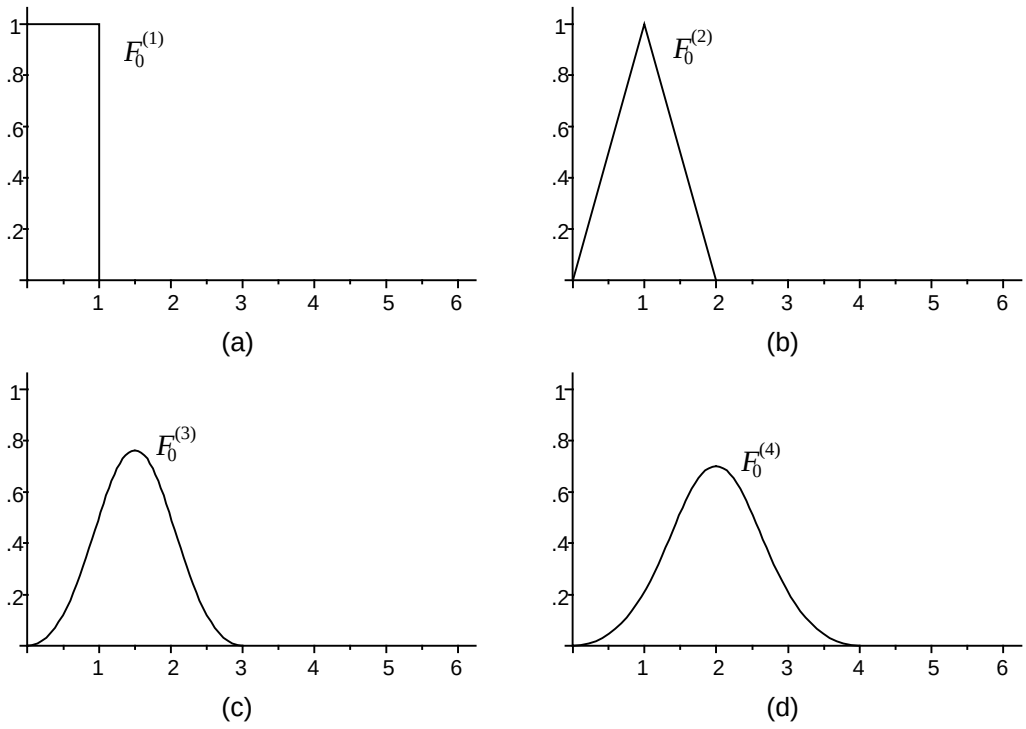


Figura A-10 Funções *B-spline*: (a) *B-spline* de primeira ordem; (b) *B-spline* de segunda ordem; (c) *B-spline* quadrática; (d) *B-spline* cúbica.

Na fig.A-10 encontram-se representadas as quatro funções *B-spline* de menor ordem. As funções de base *B-spline* têm um suporte finito[5]; as de ordem superior a um têm apenas um máximo no intervalo onde estão definidas; nunca são negativas; e a soma de todas as *B-splines* da mesma ordem é constante e igual a um. Para além destes aspectos, a principal propriedade das *B-splines* de ordem *r* é servirem de base ao subspaço de todas as funções que respeitem as seguintes características [Unser (91)]:

- serem curvas segmentalmente polinomiais de grau $r-1$;
- serem contínuas em toda a sua extensão;
- terem igualmente as suas derivadas de ordem 1, 2, ..., $r-2$ contínuas em toda a sua extensão.

Portanto, qualquer função $\varphi_r(t)$ pertencente ao subspaço referido tem continuidade C^{r-2} e pode ser expressa através da seguinte relação,

[5] A função $F_k^{(r)}(t)$ anula-se fora do intervalo $\left[t_k, t_{k+r}\right[$.

$$\varphi_r(t) = \sum_{k=-\infty}^{+\infty} a_k F_k^{(r)}(t). \qquad \text{(A-37)}$$

Se, por exemplo, as funções de base em (A-37) forem funções polinomiais de quarta ordem, a função resultante $\varphi_4(t)$, correspondendo a uma curva segmentalmente polinomial de ordem quatro, é contínua em C^2.

A função $\varphi_r(t)$ é univocamente determinada pelos seus coeficientes *B-spline* $\{a_k\}$. Estes coeficientes podem representar as amostras de um determinado sinal ou então corresponderem a meros pontos de controle. É precisamente a suavidade destas funções e o suporte finito das funções de base que tornam as *B-splines* atractivas. Para além das propriedades já mencionadas, as *B-splines* apresentam ainda as seguintes características:

- as funções de base são reais;
- a curva polinomial tende a seguir a forma do polígono de construção, logo não passa necessariamente pelos pontos de controle ou amostras;
- a curva encontra-se totalmente inserida no interior do maior polígono convexo que é possível obter a partir dos vértices correspondentes aos pontos de controle utilizados;
- a curva nunca oscila em torno de qualquer linha recta mais vezes que o respectivo polígono de construção (polígono com os vértices dados pelos pontos de controle da curva);
- pode ser introduzida qualquer alteração pontual na curva mexendo apenas em um ou mais pontos de controle.

Note-se que, desde que se tenha uma frequência de amostragem constante, todas as *B-splines* com a mesma ordem *r* têm a mesma forma, sendo cada uma delas apenas uma versão deslocada de qualquer outra. Podemos então identificar todas as funções *B-spline* com a mesma ordem *r* a partir de apenas de uma delas, que poderá ser, por exemplo, a primeira. Assumindo um período de amostragem unitário (T=1), esta função pode ser obtida recursivamente através do operador de convolução, "$*$", da seguinte forma [Schoenberg (73)]

$$F_0^{(1)}(t) = \begin{cases} 1 & \text{se } 0 \le t < 1 \\ 0 & \text{caso contrario} \end{cases}, \text{ e} \tag{A-38a}$$

$$F_0^{(r)}(t) = F_0^{(r-1)}(t) * F_0^{(1)} \quad \text{para} \quad r = 2,3,\ldots \tag{A-38b}$$

É possível verificar este tipo de recursividade através da fig.A-10. Note-se que, à excepção da primeira, cada uma das funções pode ser obtida por convolução da função imediatamente anterior com a de ordem mais baixa, tal como expresso nas equações (A-38).

Interpretando a expressão de uma forma não recursiva, conclui-se que uma *B-spline* de ordem *r* pode ser obtida com a convolução sucessiva de *r* funções rectangulares $F_0^{(1)}$ dadas por (A-38a). Ou seja,

$$F_0^{(r)}(t) = \overbrace{F_0^{(1)}(t) * F_0^{(1)}(t) * \cdots * F_0^{(1)}(t)}^{r\times} \tag{A-39}$$

E como já mencionado, qualquer *B-spline* de ordem *r* é obtida deslocando a função $F_0^{(r)}(t)$, ou seja,

$$F_k^{(r)}(t) = F_0^{(r)}(t-k)\,. \tag{A-40}$$

Uma vez que as funções se encontram deslocadas umas das outras com deslocamentos iguais ao período de amostragem e a duração de cada função é igual a esse deslocamento multiplicado pela sua ordem (ou seja $r \times T$), as funções encontram-se com sobreposições de $(r-1)/r$ %. Deste modo, as *B-splines* quadráticas, por exemplo, possuem sobreposições de $2/3$, tal como ilustrado na fig.A-11.

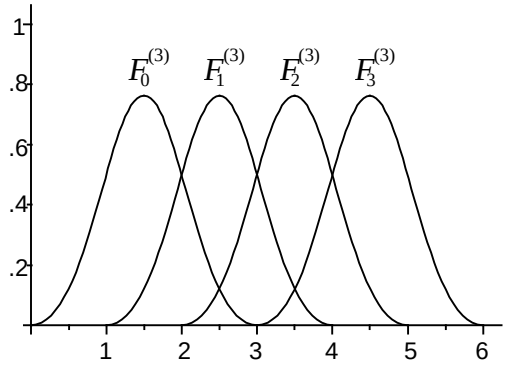


Figura A-11 Funções *B-spline* quadráticas.

A função *B-spline* contínua de ordem *r* encontra-se também definida por [Unser (91)] da seguinte forma,

$$F_0^{(r)}(t) = \sum_{j=0}^{r} \frac{(-1)^j}{(r-1)!} \binom{r}{j} (t-j)^{r-1} \mu(t-j), \tag{A-41}$$

$$\text{com} \quad \mu(t) = \begin{cases} 1 & \text{para } t \geq 0; \\ 0 & \text{para } t < 0; \end{cases} \quad \text{e} \quad \binom{r}{j} = \frac{r!}{(r-j)!\, j!}.$$

Tal como acontecia com as funções *Bézier*, se dispusermos de um polígono de construção, de vértices $B_0, B_1, \cdots, B_n$, podemos expressar a curva *B-spline* resultante, $P(t)$, como uma soma ponderada desses vértices, com os respectivos pesos definidos pelas funções de base *B-spline*, respectivamente $F_0^{(r)}, F_1^{(r)}, \cdots, F_n^{(r)}$. Uma vez que os pontos de controle poderão representar amostras de um sinal a reconstruir, as funções devem sofrer um avanço de $r/2$, de modo a ficarem centradas em relação a cada uma das amostras ou pontos de controle. Desta forma, a curva será dada por

$$P(t) = \sum_{k=0}^{n} B_k F_k^{(r)}(t + r/2). \tag{A-42}$$

Uma vez que as funções de base *B-splines* se encontram sobrepostas e não existem funções antes do primeiro e depois do último pontos de controle, as secções de curva correspondentes às zonas de influência destes pontos de controle, B_0 e B_n, não se encontram totalmente definidas. Sendo a sobreposição de $(r-1)/r$ %, apenas $1/r$ % dessas zonas se encontram definidas. Assim, se considerarmos t_0 o instante referente ao primeiro ponto de controle, e t_n o instante referente ao último ponto de controle, a curva *B-spline* encontra-se definida apenas entre $t_0 + r/2 - 1$ e $t_n - r/2 + 1$. Por exemplo, para *B-splines* de terceira ordem, a curva resultante vai desde $t = t_0 + 1/2$ até $t = t_n - 1/2$, tal como ilustrado na fig.A-12.

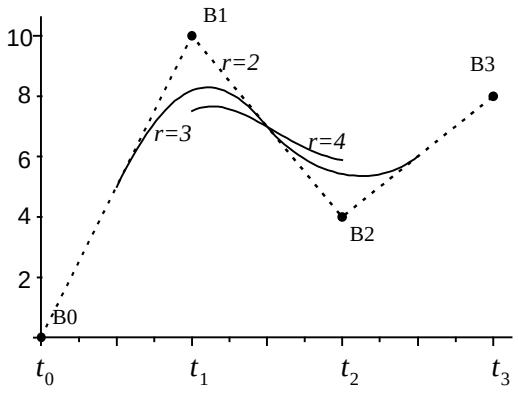


Figura A-12 Influência da ordem *r* na forma das curvas *B-spline*.

Esta figura mostra três curvas *B-splines* de diferentes ordens. Cada uma das curvas é definida através das funções de base de uma ordem específica, mas sempre a partir dos mesmos pontos de controle. Note-se que para as *B-splines* de maior ordem, as curvas encontram-se definidas num menor intervalo. É possível também verificar que a curva da *B-spline* de segunda ordem coincide com o polígono de construção e que o aumento da ordem conduz a curvas mais suaves e por isso a maior afastamento da curva *B-spline* aos pontos de controle.

Até aqui, toda a teoria apresentada sobre *B-splines* supôs a utilização da técnica *"curve fairing"*. A curva *B-spline* era gerada a partir de um conjunto de pontos de controle conhecidos que formavam os vértices do polígono de construção. Desse modo, a curva limitava-se a ter uma forma relacionada com o polígono de construção, não passando necessariamente por cada um dos pontos de controle. Existe, no entanto, também a possibilidade de utilização das *B-splines* segundo uma técnica de *"curve fitting"*. O procedimento a seguir consiste em determinar os vértices do polígono que gere uma curva *B-spline* que passe o mais próximo possível por uma série de amostras conhecidas. Este tipo de problema encontra-se ilustrado na fig.A-13.

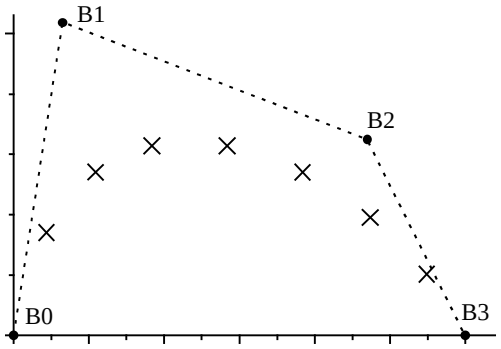


Figura A-13 Determinação do polígono de construção, dado um conjunto de amostras.

Como se pretende que as amostras se situem em cima da curva *B-spline*, a função da equação (A-42) terá que verificar as seguintes igualdades

$$D_j = P(t_j) = \sum_{k=0}^{n} B_k F_k^{(r)}(t_j + r/2) \quad para \quad j = 0,1,\cdots m \tag{A-43}$$

onde (t_j, D_j) representa as coordenadas da j ésima amostra, $m+1$ o número de amostras utilizadas e $2 \le r \le n+1 \le m+1$. Desenvolvendo a equação (A-43) para cada amostra, vem

$$\begin{array}{lcl} D_0 & = & B_0 F_0^{(r)}(t_0 + r/2) + B_1 F_1^{(r)}(t_0 + r/2) + \cdots + B_n F_n^{(r)}(t_0 + r/2) \\ D_1 & = & B_0 F_0^{(r)}(t_1 + r/2) + B_1 F_1^{(r)}(t_1 + r/2) + \cdots + B_n F_n^{(r)}(t_1 + r/2) \\ \vdots & & \\ D_m & = & B_0 F_0^{(r)}(t_m + r/2) + B_1 F_1^{(r)}(t_1 + r/2) + \cdots + B_n F_n^{(r)}(t_1 + r/2) \end{array} \tag{A-44}$$

Passando para a forma matricial, fica

$$\begin{bmatrix} D_0 \\ D_1 \\ \vdots \\ D_m \end{bmatrix} = \begin{bmatrix} F_0^{(r)}(t_0 + r/2) & F_1^{(r)}(t_0 + r/2) & \cdots & F_n^{(r)}(t_0 + r/2) \\ F_0^{(r)}(t_1 + r/2) & F_1^{(r)}(t_1 + r/2) & \cdots & F_n^{(r)}(t_1 + r/2) \\ \vdots & \vdots & \ddots & \vdots \\ F_0^{(r)}(t_m + r/2) & F_1^{(r)}(t_m + r/2) & \cdots & F_n^{(r)}(t_m + r/2) \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_n \end{bmatrix} \tag{A-45a}$$

ou

$$\mathbf{D} = \mathbf{F} \times \mathbf{B} \,. \tag{A-45b}$$

Se o número de amostras for igual ao número de funções *B-spline* utilizadas ($n = m$) a matriz **F** é quadrada e os vértices do polígono de construção são directamente obtidos por inversão da matriz, ou seja

$$\mathbf{B} = \mathbf{F}^{-1} \times \mathbf{D} \,. \tag{A-46}$$

Neste caso a curva resultante passa exactamente por todas as amostras pretendidas, tal com requerido na técnica de *"curve fitting"*.

Se, por outro lado, for desejável aumentar a suavidade da curva, minimizando dessa forma as ondulações, ou então se se pretender reduzir o número de funções *B-spline* utilizadas, de modo a minimizar quantidade de parâmetros que descrevem a curva (para efeitos de compactação de dados, por exemplo), pode-se especificar um polígono de construção com o número de vértices menor do que o número de amostras, isto é, $n < m$. Neste caso a matriz **F** passa a não ser quadrada, não sendo, por isso, possível resolver o sistema (A-44). É, no entanto, possível transformar **F** numa matriz quadrada e dessa forma poder ser invertida para se obter os vértices que determinam a curva *B-spline*. Contudo, agora a curva já não passa nas amostras como pretendido no sistema inicial, mas sim nas suas proximidades.

Como o produto de uma matriz pela sua transposta resulta sempre numa matriz quadrada, os vértices do polígono de construção pretendido são obtidos da seguinte forma

$$\mathbf{D} = \mathbf{F} \times \mathbf{B}$$

$$\mathbf{F}^T \times \mathbf{D} = \mathbf{F}^T \times \mathbf{F} \times \mathbf{B}$$

$$\mathbf{B} = \left[\mathbf{F}^T \times \mathbf{F}\right]^{-1} \times \mathbf{F}^T \times \mathbf{D} \qquad \text{(A-47)}$$

Na fig.A-14 encontram-se ilustradas as duas versões deste tipo de técnica. Enquanto que na fig.A-14a é utilizado um número de vértices igual ao número de amostras, na fig.A-14b o número de vértices é menor, por isso a curva não passa pelas amostras.

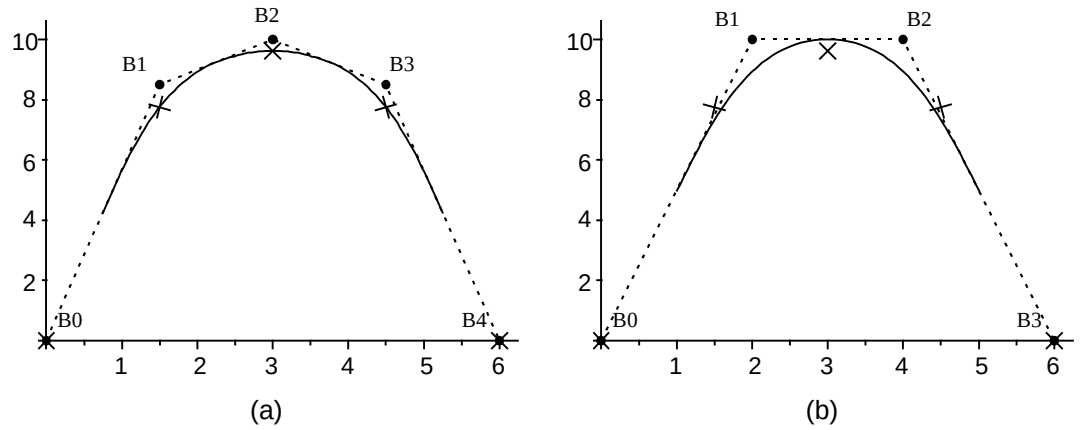


Figura A-14 Determinação de polígonos de construção, dado um conjunto de 5 amostras: (a) Polígono de 5 vértices; (b) Polígono de 4 vértices.

Referências Bibliográficas

[Aldroubi (92)] Aldroubi A., Unser M., and Eden M. (92), "Cardinal Spline filters: Stability and convergence to the ideal sinc interpolator", Signal Processing, 28:127–138, 1992.

[Alengrin (86)] Alengrin G., Barlaud M., and Menez J. (86), "Unbiased parameter estimation of nonstationary signals in noise", *IEEE* Transactions on Acoustics, Speech, and Signal Processing, 34(5):1319–1322, October 1986.

[Almeida (84)] Almeida L., e Silva F. (84), "Variable-frequency synthesis: An improved harmonc coding scheme", *IEEE ICASSP-84*, San Diego, CA, p. 27.5.1, 1984.

[Amir (89)] Amir N., and Gath I. (89), "Segmentation of EEG during sleep using time –varying autoregressive modeling", *Biological Cybernetics*, 61:447–455, 1989.

[Atal (82)] Atal B., and Remde J. (82), "A new model for LPC excitation for producion natural sounding speech at low bit rates", *Proc. ICASSP-82*, pp. 614-617, Apr. 1982.

[Atal (89)] Atal B., Cox R., and Kroon P. (89), "Spectral quantization and interpolation for CELP coders ", *Proc. ICASSP-89*, pp. 69-72, 1989.

[Campbell (86)] Campbell J., tremain T., and Welch V. (86), "Voiced/unvoiced classification of speech with applications of U.S. Government LPC-10e algorithm", *Proc. ICASSP-86*, pp. 473-476, 1986.

[Charbonnier (87)] Charbonnier R., Barlaud M., Alengrin G., and Menez J. (87), "Results of AR–modeling of nonstatinary signals", *Signal Processing*, 12:143–151, 1987.

[Chung (89)] Chung J., and Schafer R., (89), "A 4.8 Kbps homomorphic vocoder using analysis-by-synthesis excitation analysis", *Proc. ICASSP-89*, p. 144, 1989.

[Correia (90)] Correia Paulo J. O. D., e Marques Pedro J. S. L. G., "Codificação de Voz", *Projecto de Licenciatura, Aveiro Setembro de 1990.*

[Crosmer (85)] Crosmer J., and Barnwell T. (85), "A low bit rate segment vocoder based on line spectrum pairs", *Proc. ICASSP-85, Tampa, FL,* pp. 240-243, Mar. 1985.

[Engels (88)] Engels W., Stark E. L., and Vogt L. (88), "On the application of an optimal Spline sampling theorem", *Signal Processing*, 14:225–236, 1988.

[Fant (60)] Fant G. (60), "Acoustic Theory of Speech Production", *Gravenhage, The Netherlands: Mounton and Co.*,1960.

[Flanagan (66)] Flanagan J. and Golden (66), "Phase vocoder", *Bell Syst. Tech. J.*, vol. 45, p. 1493, Nov. 1966.

[FS-1015 (84)] Federal Standarrd 1015 (84), "Telecommunications: Analog to digital cnversion of ratio voice by 2400 bit/s lnear preictive coding, national communication system", *Natinal Communication System - Office of Technology and Standards*, Nov. 1984.

[Gersho (83)] Gersho A. and Cuperman V. (83), "Vector Quantization: A pattern matching technique for speech coding", *IEEE Commun. Mag.*, vol. 21, p. 15, Dec. 1983.

[Grenier (91)] Grenier Yves (91), "Time–dependent ARMA modeling of nonstationary signals", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(4):899–911, August 1983.

[Hall (83)] Hall Mark G., Oppenheim Alan V., and Willsky Alan S. (83), "Time–Varying Parametric Modeling of Speech", *Signal Processing*, 5:267–285, 1983.

[Hardwick (88)] Hardwick J., and Lim J. (88), "A 4800 bps multiband excitation speech coder", *Proc. ICASSP-88,* pp. 374-377, Apr. 1988.

[Hardwick (91)] Hardwick J., and Lim J. (91), "The application of the IMBE speech coder to mobile communications", *Proc. ICASSP-91*, pp. 249-252, May. 1991.

[Haykin (91)] Haykin Simon (91), "Adaptative Filter Theory", *Englewood Cliffs, NJ: Prentice-Hall, second edition*, 1991.

[Higdon (67)] Higdon A., Ohlsen E., Stiles W., and Weese J. (67), "Mechanics of Materials", *John Wiley & Sons, second edition*, New York, 1967.

[Holmes (80)] Holmes J. N. (80), "The JSRU channel vocoder", *Proc. Inst. Elec. Eng.* vol. 127, pt F, no. 1, pp 53-60, Feb. 1980.

[Jayant (84)] Jayant N. S., and Noll P. (84), "Digital Coding of Waveforms", *Englewood Cliffs, NJ: Prentice-Hall*, 1984.

[Louis (75)] Louis A. Liporace (75), "Linear estimation of nonstationary signals", *The Journal of the Acoustical Society of America*, 58(6):1288–1295, December 1975.

[Magill (75)] Magill D., and Un C. (75), "The residual-excited linear prediction vocoder with transmission rate below 9.6 Kbits/s", *IEEE Transactions Commun.*, vol. COM-23, no. 12, p. 1466, Dec. 1975.

[Makhoul (75)] Makhoul J. (75), "Linear prediction: A tutorial review", *Proc. IEEE*, vol. 63, no. 4, pp. 561-580, Apr. 1975.

[Marques (90)] Marques J., Almeida L., e Tribolet J. (90), "Harmonic coding at 4.8 kb/s", *Proc. ICASSP-90*, New Mexico, p. 17, Apr. 1990.

[McAulay (86)] McAulay R., and Quatieri T. (86), "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol ASSP-34, no. 4, p. 744, August 1986.

[McCree (93)] McCree A., and Barnwell III. (93), "Implementation and evaluation of a 2400 bps mixed excitation LPC vocoder", *Proc. ICASSP-93*, Minneapolis, p. II-159, Apr. 1993.

[Portnoff (73)] Portnoff M. R., Schafer R. W. (73), "Mathematical Considerations in Digital Simulations of the Vocal Tract", *The Journal of the Acoustical Society of America*, 53(1), p. 294, January 1973.

[Rabiner (79)] Rabiner L. R. and Schafer R. W. (79), "Digital Processing of Speech Signals", *Prentice-Hall Signal Processing Series*, 1979.

[Ries (91)] Ries Sigmar (91), "On the reconstruction of signals by a finite number of samples", *Signal Processing*, 23:45–68, 1991.

[Rogers (90)] Rogers David F. and Adams J. Alan (90), "Mathematical Elements for Computer Graphics", *McGraw-Hill, second edition*, 1990.

[Schoenberg (73)] Schoenberg I. J. (73), "Cardinal Spline Interpolation", *Regional Conference Series in Applied Mathematics*, volume 12, SIAM, Philadelphia, 1973.

[Schroeder (84)] Schroeder M., and Atal B. (84), "Code-Excited Linear Prediction (CELP): Hight-Quality Speech at Very Low Bit Rates", *Proc. Int. Conf. Commu.*, pp. 1610-1613, 1984.

[Singhal (84)] Singhal S., and Atal B. (84), "Improving the performance of multi-pulse coders at low bit rates", *Proc. ICASSP-84*, p. 131, 1984.

[Silva (93)] Silva Tomás O. (93), "On the Application of an Optimal Spline Sampling Theorem to Parametric Modeling of Nonstationary Signals", *Universidade de Aveiro, Portugal, 1993*.

[Sondhi (74)] Sondhi M. M. (74), "Model for Wave Propagation in a Lossy Vocal Tract", *The Journal of the Acoustical Society of America*, 55(5), pp. 1070–175, May 1974.

[Spanias (94)] Spanias Andreas S. (94), "Speech Coding: A Tutorial Review", *Proceedings of the IEEE*, 82(10):1541–1582, October 1994.

[Trancoso (87)] Trancoso Isabel Maria M. "High Quality Speech Coding at Medium-to-low Bits Rates", *Tese de Doutoramento, Lisboa Março de 1987*.

[Tremain (93)] Tremain Thomas, Kemp David, Collura John, and Kohler Mary (92), "Evaluation of Low Rate Speech Coders for HF",*U. S. Department of Defense, Proc. ICASSP-93,* April 1993.

[Unser (91)] Unser Michael, Aldroubi Akram, and Eden Murray (91), "Fast B-spline transforms for continuous image representation and interpolation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):277–285, March 1991.

[Vary (88)] Vary P., (88), "Speech codec for the european mobile radio system", *Proc. ICASSP-88,* p. 277, Apr. 1988.

[Zelinski (77)] Zelinki R. and Noll P. (77), "Adaptive transform coding of speech signals", *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-25, p. 299, Apr. 1977.